National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file   Votre référence

Our file   Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

# DEVELOPMENT OF AN INSTRUCTIONAL COMPONENT FOR THE PREPARATION OF TEACHERS IN CLASSROOM ASSESSMENT

BY

HENRY W. SCHULZ ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

SPRING, 1994

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file   Votre référence

Our file   Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-612-11362-0

Canada

## UNIVERSITY OF ALBERTA

## RELEASE FORM

**NAME OF AUTHOR:** Henry W. Schulz

**TITLE OF THESIS:** Development of an Instructional Component for the Preparation of Teachers in Classroom Assessment

**DEGREE:** Doctorate in Educational Psychology

**YEAR THIS DEGREE GRANTED:** 1994

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis, and to lend or sell such copies for private, scholarly or scientific purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

207 Waterford Bridge Road
St. John's, Newfoundland
A1E 1E4

November 26, 1993

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

THE UNDERSIGNED CERTIFY THAT THEY HAVE READ, AND RECOMMEND TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH FOR ACCEPTANCE, A THESIS ENTITLED **DEVELOPMENT OF AN INSTRUCTIONAL COMPONENT FOR THE PREPARATION OF TEACHERS IN CLASSROOM ASSESSMENT** SUBMITTED BY HENRY W. SCHULZ IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN EDUCATION IN EDUCATIONAL PSYCHOLOGY.

Maguire
Dr. T. Maguire, supervisor

S. Hunka
Dr. S. Hunka

T. Rogers
Dr. T. Rogers

Dr. L. Wankel

Dr. L. Beauchamp

R. Wilson
Dr. R. Wilson

November 24, 1993

# DEDICATION

To my mother who, unfortunately, did not live to see this day.

To Anita who understands what this means, to Bretta and Stefan who understand only as children can, and who do live to see this day.

# ABSTRACT

The purpose of the study was to develop an instructional component for preparing teachers in classroom assessment. To provide a basis for what is realistically possible in classroom assessment, four exemplary teachers were selected for intensive study: two taught junior high science and two taught junior high social studies. The assessment practices of these teachers were observed over a period of time, and their assessment documents for this time period were collected. They were then interviewed regarding assessment purposes and practices.

From an analysis of these case studies, and the literature review, six teacher competencies were outlined under two broad categories: Using assessment information effectively, and Designing and conducting appropriate assessments.

Four characteristics of classroom assessment were identified from measurement and evaluation theory: reliability, validity, utility, and efficiency. These were used as the framework for developing 17 recommendations for teacher preparation in assessment. The recommendations were critically reviewed by 16 educators with backgrounds including curriculum, assessment, teacher supervision, and teaching.

Based on the review and the competencies derived from the case studies, it was finally recommended that teachers receive instruction in:

1. Determining what is to be assessed, and the most appropriate ways to assess these learnings. The instruction should integrate principles of measurement, cognitive development, and curriculum and teaching.

2. Designing assessments which provide information for formative and summative purposes. This includes designing assessments that are appropriately weighted in content and skill level, and that provide usable scores.

3. Producing good paper and pencil assessments relevant to learning goals.

4. Developing procedures for, and directly assessing, students' products and behaviours.

5. Identifying important affective objectives, and in how to assess them meaningfully and fairly. These arise in subject areas and from societal values more generally.

6. Combining information from assessments and interpreting this information to students, parents, and others.

7 The qualities of good assessment procedures, and the need for obtaining sufficient information to make important decisions.

8. Theoretical bases of measurement and evaluation, with emphasis on validity, including bias and subjectivity, and how this impinges on classroom practice.

# ACKNOWLEDGEMENTS

I would like to acknowledge those who assisted me in completing this study. To Dr. Tom Maguire a special thanks for the time and support he gave, and the patience he offered in bringing this to completion. To Dr. Steve Hunka also a special thanks for his willing assistance. I will always value both of you for your wisdom.

I also appreciate the effort of members of the committee, and Dr. Todd Rogers in particular for his comments and advice.

The assistance of the teachers who participated in the case studies, and the educators who gave of their time to review my recommendations, is greatly appreciated. Without this willing effort my research would not have been possible.

Finally, I wish to acknowledge and thank Anita for her support, and my children who, for their whole lives to this day, accepted a father beleaguered by a task that seemed insurmountable.

## TABLE OF CONTENTS

I. INTRODUCTION .................................................. 1

**Nature of the Study** .................................................. 2

**Organization of the Dissertation** .................................................. 3

II. REVIEW OF THE LITERATURE .................................................. 4

**Background to Classroom Assessment Practices** .................................................. 6

Teachers and Standardized Test Use .................................................. 6

Comments on Standardized Test Use in the Classroom .................................................. 11

**Teachers' Classroom Assessment Practices** .................................................. 12

In-class Assessment Time .................................................. 12

Classroom Assessment Purposes and Practices .................................................. 13

Describing classroom assessment .................................................. 14

Teacher assessment practices .................................................. 17

Teacher grading practices .................................................. 22

Quality of Teacher-Made Assessments .................................................. 25

Summary of Classroom Assessment Practices .................................................. 27

**Teacher Training in Educational Assessment** .................................................. 28

Content Covered in Measurement Textbooks .................................................. 31

Course Content for Teacher Preparation in Educational Assessment .................................................. 32

**Summary** .................................................. 36

**Purpose and Focus of the Study** .................................................. 36

Purposes of the Study .................................................. 37

Focus and Features of the Study .................................................. 37

III. THE CASE STUDIES .................................................. 39

**Organization of the Chapter** .................................................. 39

**Selection of Teachers for the Case Studies** .................................................. 39

Rationale for the Teachers to be Studied .................................................. 40

Characteristics of the Teachers Selected .................................................. 41

**Case Study Procedures** .................................................. 41

**Classroom Observations** .............................................. **42**

**Teacher Interviews** .................................................. **43**

**Teacher Interviews--Structured Responses to Six Aspects of Assessment** .................................................. **44**

1. Importance of Various Purposes ........................................ 44
2. Importance of Various Methods to Assess Achievement .............. 45
3. Importance of Various Methods to Assess Affect ...................... 46
4. Importance of Various Criteria for Selection of Assessment Methods ........................................................ 47
5. The Importance of Various Sources of Assessment Knowledge ........................................................ 48
6. Manner in which Teachers Allocate Their Assessment Time ......... 49

**Teacher Interviews--Unstructured Responses and Observations** .................................................. **50**

1. Assessment Purposes ...................................................... 50
    - Diagnosing individual needs of students ............................ 50
    - Diagnosing group needs of students .................................. 51
    - Assigning grades ........................................................ 51
    - Grouping for instruction ............................................... 52
    - Identifying students for special services ............................ 53
    - Controlling and motivating students .................................. 53
    - Evaluating instruction ................................................... 53
    - Communicating achievement expectations ........................... 54
    - Communicating affective or behavioral expectations ............... 54
    - Providing test taking experience ...................................... 55
2. Methods Used by Teachers for Assessing Achievement .............. 55
    - Teacher-developed paper and pencil tests and quizzes .............. 55
    - Text-embedded paper and pencil tests and quizzes .................. 56
    - Performance assessment ................................................ 56
    - Oral questioning strategies, interviews, and presentations ......... 57
    - Standardized tests ....................................................... 58
    - Group assessment methods ............................................. 58
    - Opinions of other teachers ............................................. 60
    - Assessment of reasoning skills ........................................ 60
    - Regular homework assignments ........................................ 60
    - Student peer ratings ..................................................... 61

Student self ratings........................................................ 61
Strategies for integrating assessment and instruction............... 61
Dealing with cheating................................................... 62
3. Methods Used by Teachers for Assessing Affect........................ 63
Observing individual students........................................... 63
Observing group interactions........................................... 64
Using questionnaires..................................................... 64
Using interviews.......................................................... 64
Opinions of other teachers regarding affective characteristics...................................................... 64
Other sources of information for affective characteristics.......... 65
Assessment of Skills by Teachers......................................... 65
4. Factors and Criteria Used in Deriving Assessment Plans............. 66
Assessment results fit purpose.......................................... 66
Assessment methods match intended learning outcomes........... 67
Less important criteria for the selection of assessment methods........................................................... 68
Criteria considered unimportant for the selection of assessment methods........................................... 69

**Quality of Classroom Assessments** ................................ **70**
Paper and Pencil Assessments............................................ 70
Quality of the test document............................................ 71
Item formats included in the tests....................................... 74
Cognitive levels required for responding to the test items.......... 76
Technical Quality of the Test Items...................................... 80
Choice format items....................................................... 81
Short-answer formats..................................................... 82
Essay formats.............................................................. 83
Context-dependent formats............................................... 83
Concluding comments..................................................... 84
Performance Assessment.................................................. 84
Assessment of processes and behaviours................................ 85
Assessment of student products.......................................... 86
Concluding comments..................................................... 87
Oral Questioning by Teachers............................................ 88
Oral questions asked by the teacher during instruction............. 88
Oral testing................................................................ 89

Summary .......... 89
Oral Feedback to Students .......... 89
Written Feedback to Students .......... 90
Reporting Student Progress .......... 90

**The Practices of Classroom Assessment and What this Means for Teacher Preparation .......... 91**

**IV. COMPONENTS OF A PROGRAM FOR TEACHER DEVELOPMENT IN CLASSROOM ASSESSMENT .......... 93**

**Assessment in the Classroom Context .......... 93**
Purposes of Classroom Assessment .......... 93
The Context of Classroom Assessment .......... 95

**Characteristics of Good Classroom Assessment .......... 96**
Reliability and Error of Measurement .......... 97
Recommendations for reliability .......... 100
Validity .......... 105
Substantive component of validity .......... 106
Structural component of validity .......... 109
External component of validity .......... 110
Recommendations for validation of classroom assessments .......... 112
Utility .......... 118
Referential basis for interpreting assessment scores .......... 119
Applicability of assessment results .......... 120
Communicability of assessment results .......... 121
Objectivity in scoring .......... 121
Recommendations for utility of classroom assessments .......... 122
Efficiency .......... 123
Recommendations for the efficiency of classroom assessments .......... 123

**Summary of Recommendations for Teacher Preparation in Classroom Assessment .......... 124**
Reliability--Recommendations 1 to 5 .......... 124
Validity--Recommendations 6 to 12 .......... 126
Utility--Recommendations 13 to 16 .......... 128
Efficiency--Recommendation 17 .......... 129

V. EDUCATOR REVIEW OF THE RECOMMENDATIONS FOR TEACHER PREPARATION IN CLASSROOM ASSESSMENT .......... 130

**Basis for Reviewing the Recommendations for Teacher Preparation in Classroom Assessment** .......... **130**

**Procedures to Review the Recommendations** .......... **133**

Individuals Selected to Review the Recommendations .......... 133

Features for Review of the Recommendations .......... 134

Structured procedures for the review .......... 134

Analysis of the responses to the structured procedures .......... 136

**Review of the Recommendations** .......... **137**

Ratings of Importance of the Recommendations .......... 138

Responses to Feature 1--Level of Specificity for Instruction .......... 140

Responses to Feature 2--Common or Differentiated Instruction .......... 141

Responses to Feature 3--Method of Delivery of Instruction .......... 143

Responses to Feature 4--Nature of Instruction .......... 144

**Implementation of the Recommendations** .......... **144**

Recommendations 7, 6, and 15 .......... 145

Recommendations 1, 2, and 10 .......... 148

Recommendations 12, 14, and 5 .......... 151

Recommendations Rated Lower in Importance .......... 153

**Implications for Instruction in Classroom Assessment** .......... **156**

Focus of the Instruction .......... 156

Approaches to Instruction .......... 158

VI. GUIDELINES FOR TEACHER PREPARATION IN CLASSROOM ASSESSMENT .......... 161

**General Findings and Implications From the Case Studies** .......... **161**

Limitations of the Case Studies .......... 161

Teacher Competencies in Classroom Assessment .......... 162

Using Assessment Information Effectively .......... 162

Designing and Conducting Appropriate Assessments .......... 166

**Specification and Review of the Recommendations** .......... **170**

Limitations of the Review of the Recommendations .......... 171

Reliability .................................................................. 171
Validity .................................................................. 172
Utility and Efficiency .................................................................. 174

**Conclusions and Recommendations for Teacher Preparation** .................................................. **176**

**Implications for Further Research and Practice** .................. **178**

**REFERENCES** .................................................................. **180**

**APPPENDICES** .................................................................. **199**

**A. Letters Related to the Case Studies** ................................ **200**

**B. Forms Used to Obtain Teachers' Structured Responses to Each of the Six Aspects Related to Classroom Assessment** .................................................... **205**

**C. Ratings of Test Item Quality in Teacher-Made Tests** .............. **212**

**D. Document Presented to Reviewers of the Recommendations** .................................................. **216**

**E. Results of the Review of the 17 Recommendations** ................ **249**

# LIST OF TABLES

Page

1. Percentage of Teachers Reporting Use of Standardized Achievement Test Results in Their Classrooms .......... 7
2. Percentages of Teachers Reporting Use of Various Assessment Techniques .......... 19
3. Relative Emphasis Accorded Major Types of Classroom Assessment by Teachers and by Six Commonly Used Measurement Textbooks .......... 32
4. Relative Emphasis Given to Various Topics by Eight Measurement Textbooks .......... 33
5. Relative Importance to the Teachers of Ten Purposes .......... 44
6. Relative Importance to the Teachers of Eleven Achievement Assessment Methods .......... 46
7. Relative Importance to Three of the Teachers of Eight Affective Assessment Methods .......... 47
8. Relative Importance to the Teachers of Nine Criteria for Selection of Assessment Method .......... 48
9. Relative Importance to the Teachers of Six Sources of Assessment Knowledge .......... 49
10. Relative Amount of Time Teachers Give to Seven Tasks in Assessment .......... 50
11. Rating of Test Documents as to Format and Directions .......... 72
12. Types of Assessment Methods Used by Two Science and Two Social Studies Teachers .......... 75
13. Marks Awarded for Item Types at Various Cognitive Levels .......... 78

## LIST OF FIGURES

Page

1. Typical Scheme Teachers Used for Grading Students on a Unit of Instruction ........ 52
2. Form for Reviewers' Structured Responses to Each Recommendation .......... 135
3. Summary Statements of the 17 Recommendations in the Order They Were Presented to Reviewers ........ 138
4. Recommendations Ordered According to Reviewers' Mean Ratings of Importance ........ 139
5. Numbers of Reviewer Responses to Levels of Specificity for Instruction in Assessment ........ 141
6. Numbers of Reviewer Responses to Level of Program Differentiation for Instruction in Assessment ........ 142
7. Summary Statements of the 17 Recommendations Ordered and Grouped According to Reviewers' Ratings of Importance ........ 145

# I. INTRODUCTION

The purpose of the study was to develop an instructional component for the preparation and development of teachers in the area of classroom assessment. This instruction was to be consistent with evaluation theory and recommended practices in classroom assessment, and was to encompass the goals and objectives of the schools. The component was intended also to reflect the realities faced by teachers in their classrooms since so much of the development in measurement theory has not translated into good assessment practices by our teachers.

Many educators have expressed concern about the quality of assessments carried out in classrooms. This is true in Canada (e.g., McLean, 1985; Rogers, 1991) as well as in the U.S. (e.g., Nitko, 1991a; Stiggins, 1988a). It appears that teachers do not generally produce good-quality tests (e.g., Chambers, 1982; Haertel, 1986; Stiggins, 1988b), nor do they have a good grasp of the technical aspects of measurement (e.g., Boothroyd, McMorris, & Pruzek, 1992; Newman & Stallings, 1982). Further, educators have maintained that teachers often do not assess the kinds of learning considered most important to our students, such as complex and sustained reasoning skills and critical thinking (e.g., Stiggins, Griswold, & Wikelund, 1989), and that they do not assess these learnings in the most appropriate ways (e.g., Wiggins, 1989b).

Stiggins, Conklin, and Bridgeford (1986) reviewed the literature on classroom assessment practices and concluded that training in measurement was not adequate, and that the courses that were available in teacher training programs did not typically provide teachers with the education appropriate to the demands of the classroom. Gullickson and Ellwein (1985) argued that much of what was taught in measurement courses was essentially irrelevant to classroom assessment practices. This appears not to have changed much and the same concern was expressed more recently (e.g., Nitko, 1991a). Also, large numbers of teachers apparently do not have any course work in assessment, neither in the U.S. (e.g., Green & Stager, 1986; O'Sullivan & Challnick, 1991) nor in Canada (e.g., McLean, 1985; Rogers, 1990b; Webster, 1987). In fact, three quarters of the teachers surveyed by Stiggins and Bridgeford (1985) indicated concern with their classroom assessment practices. This apparent lack of adequate preparation is an old concern, and has been voiced frequently over the years: for example, Ebel (1967), Roeder (1973), and Gullickson (1982).

It is well established that teachers prefer assessment materials that they develop themselves to those prepared by outside agencies (e.g., Gullickson, 1985; McLean, 1985). Teachers make little use of any form of standardized tests, including those produced by state or provincial agencies. Teachers also are being asked to teach and assess a broad and complex array of skills, knowledge, thinking, and attitudes. These often require alternative assessment techniques for which materials are not readily available, and many are based on direct observation of student behaviours. This makes it imperative that teachers have the ability to develop the assessment procedures necessary for their programs and classrooms. One clear way for teachers to develop these abilities is through pre-service and in-service training.

Recently, there have been many developments in the area of classroom assessment. This is in marked contrast to a few years ago when the research emphasis was clearly on standardized tests and large-scale assessment programs (Stiggins, 1985). Standards have been prepared for teacher competence in assessment (American Federation of Teachers et al., 1990), principles to guide assessment practices in the classroom have been agreed upon (*Principles for fair student assessment practices for education in Canada*, 1993), and educators are attempting to define what teachers ought to be taught and how this should be done (e.g., Nitko, 1991a).

Although research has begun in the field of classroom assessment as distinct from large-scale testing, further effort is needed to clarify how teachers go about their assessment tasks, and how these can be improved. For example, it is not clear what the needs of teachers are, and what instruction is necessary to assist them in assessing students. Also, we have very little systematic information on how assessment results are communicated to students and parents, and how they are understood--or even if they are.

## Nature of the Study

The study included two major phases. The first consisted of intensive study of a limited number of exemplary teachers to determine the assessment practices that are used by good teachers, and therefore reasonably possible in classrooms. This was to provide an indication of an ideal setting, but one that retained its grounding in actual classroom practice. There is considerable survey evidence regarding general questions of classroom practices and teacher measurement knowledge, but only limited in-depth information as to what happens in the classroom in terms of assessment, and what is realistically possible. Thus, it was decided to carry out case studies of exemplary teachers using observations and interviews.

The second phase of the study involved the development and critical review of recommendations for teacher preparation and development in classroom assessment. First, recommendations were drafted with a focus on preservice education. They were based on measurement and evaluation theory, but reflected the research on classroom assessment and incorporated the interpretations of classroom practice from the case studies. The rationale underlying the recommendations and the recommendations themselves are described in Chapter IV (also summarized in Appendix D). Second, to ensure that the recommendations were in keeping with current thinking in the area and practicable for teachers in the classroom, they were submitted for critical review by a number of educators serving in various capacities. The reviewers included professors of measurement education and educational measurement specialists working in school systems, professors of curriculum and curriculum specialists in school systems, administrators with direct interests in assessment, and classroom teachers. These individuals were asked to review each of the recommendations regarding its importance for the preparation of teachers, and to indicate how this recommendation could best be brought about in preservice teacher education. The reviewers were also asked to comment generally. The procedures used for review and the results of the review are described in Chapter V. The document submitted to reviewers details of the review are presented in Appendix D.

Lastly, the results of the case studies were summarized, and from this six general competencies in assessment were identified for teachers. The critical reviews of the recommendations were then summarized, and a number of principles were specified that can guide instructional development for teacher preparation in classroom assessment. The principles were intended to guide curriculum development, but stop short of specifying content and objectives for the course. Course suggestions are readily available (e.g., Linn, 1990; Rogers, 1991), and measurement textbooks abound. Unfortunately, textbooks historically have not addressed many of the important issues identified for teacher preparation, and, for example, Stiggins and Bridgeford (1985) showed that much of textbook content was inappropriate. There have been some changes in textbooks since this analysis, and many new books as well as many new editions of older ones have appeared. Some of these books are beginning to accommodate the changes called for (e.g., Airasian, 1992), but an informal review indicated that these changes are limited to date.

## Organization of the Dissertation

In the chapter that follows, Chapter II, some of the assessment literature is reviewed. In Chapter III the case studies are described and analyzed, with the focus on teachers' classroom assessment practices (the first phase of the study). The second phase of the study is described in Chapters IV and V. Chapter IV provides a detailed rationale based on measurement theory for the recommendations for teacher preparation in classroom assessment, and presents 17 recommendations. Chapter V describes how these recommendations were reviewed by experienced educators, and indicates which recommendations were considered most important for teacher education and how these could be implemented. Finally, Chapter VI gives a synopsis of the case studies and what implications these have for assessment education, a summary of the review of the recommendations, and a set of principles for guiding instructional development in classroom assessment for prospective teachers.

## II. REVIEW OF THE LITERATURE

The importance of student evaluation in our schools is well recognized by educators and the general public alike, although educators may have a broader view of the role of evaluation in education than that evidenced by the public's concern for accountability (e.g., Airasian, 1988a; Hathaway, 1983; McLean, 1985; Rogers, 1990a, 1991). Educators typically view student evaluation as integral to the teaching-learning environment--for system and program monitoring, for instructional planning, for student motivation, and for student grading and reporting (e.g., Dorr-Bremme & Herman, 1986). The public in general appears more interested in overall indicators of system performance as provided by large-scale assessment programs, such as province- or state-mandated testing programs, or local testing programs which often employ commercial standardized achievement tests (McLean, 1985; D. P. Resnick, 1981; Kirst, 1991a). That is, representatives of the public (elected officials at various levels) often tend to view assessment in its "monitoring of the schools" role, accountability, and call for standardized achievement testing of one form or another to provide the information. Examples of this at the national level include the National Assessment of Educational Progress in the U.S. (Algina, 1992; Messick, 1985), the Assessment of Performance Unit in the United Kingdom (Burstall,1986), and the national assessment program implemented by the Council of Ministers of Education, Canada (Rogers, 1990b).

Most provinces in Canada (Jones & Carbol, 1988; Schulz, 1985) and states in the U.S. (Airasian, 1988a; Pipho, 1980; Popham, 1991; D. P. Resnick, 1980) have assessment programs of some form or other with the purposes usually of system monitoring, public reporting, maintaining standards, and providing curriculum feedback, and in many cases student evaluation. However, the assessment programs and their purposes vary considerably from one jurisdiction to another, and also from Canada to the U.S.: for example, no Canadian province had instituted a minimum-competency testing program (Jones & Carbol, 1988) whereas in 1980 more than 40 states had done so (Pipho, 1980). In 1988, six provinces (Jones & Carbol, 1988) and 29 states (Airasian, 1988a) required students to write one or more high school exit exams to receive a diploma. The trend to high-stakes testing in the U.S. appears to continue, and Popham (1991) states:

> state after state enacted laws requiring pupils to take annually administered achievement tests. The results of such tests were used not only to determine the quality of statewide schooling but also . . . to mirror the effectiveness of individual school districts and schools. In every sense of the expression, these legislatively mandated achievement tests were *high-stakes* because there were significant contingencies associated with the test results. . . . To students, the tests were significant because in many instances test scores were linked to high school graduation or grade-level promotion. To educators, because of the manner in which test results were published by the local media, high test scores were viewed as indicating an effective instructional program and, of course, vice versa. (p. 12)

Local or district-level standardized testing appears to be flourishing as well. Sproull and Zubrow (1981) state that "every U.S. school system supports a standardized testing program. . . . All school systems employ at least one achievement test. . . . most school districts use achievement tests in all grades up to high school" (pp. 628, 629). Lazar-Morrison, Polin, Moy, and Burry (1980) reported a figure of 75% of American school districts that have district-wide standardized achievement testing, and that over 80% of school students received standardized achievement testing. In an earlier study, Goslin (1967) reported that all 75 secondary school principals surveyed, and 713 of 714 elementary school principals, indicated some standardized tests were administered in their schools. Relatively few of these principals reported not administering standardized achievement tests. Kirkland (1971) stated that the Coleman study reported 90% of U.S. students as being in schools where intelligence and achievement tests were administered at both elementary and secondary levels, and concluded from test sales that students receive from three to five standardized tests per pupil per year. Standardized test sales increased dramatically during the 1960's and 1970's suggesting that this type of testing has increased in prevalence (D. P. Resnick, 1981). It is unlikely that many students in U.S. schools escape some form of commercial standardized ability or achievement testing in their school years. Neill and Medina (1989) state "A recent study by the National Center for Fair and Open Testing estimated that U.S. public schools administered 105 million standardized tests to 39.8 million students during the 1986-87 school year. . . . an average of more than 2.5 standardized tests per student per year" (p. 688).

The advent of state-mandated minimal competency and other forms of testing has increased the amount of testing originating from outside the classroom (D. P. Resnick, 1980; Pipho, 1978, 1980; Popham, 1991). In the wake of the public cry for accountability and the pressures put on schools to increase learning Hathaway (1983) argued that there is "renewed commitment to testing" (p. 1), but he noted that there is also "recognition of the primacy of the classroom as the focus for testing and test use" (p. 1). That is, despite the considerable effort given to large-scale achievement assessment programs by administrative systems, teachers in classrooms are the key users of testing and evaluation involving students. This point can be illustrated by a dramatic account of what testing takes place in schools. Yeh, Herman, and Rudner (1981), who surveyed test use in 20 California elementary schools, state:

> All schools in the study administered yearly state assessment tests in Grades one, two, three, and six, and all administered annual or semi-annual standardized norm-referenced test batteries to students within their purview. A sizeable number were required, in addition, to give beginning and end of year assessments of a criterion-referenced or district continuum variety. As with all California schools, the schools in the study were also involved in required minimum competency testing. While this listing of required tests is sizeable, it is not exhaustive, and only begins to tell the story of how much testing goes on in schools. *Other kinds of tests, teachers reported, constituted a much greater proportion of assessment activities in schools* [italics added]. (p. 2)

Is the story regarding classroom testing much different in Canada? Probably not. McLean (1985) suggests that most, but not all, school districts administer some standardized achievement and/or ability tests to students at several grade levels. Schulz (1985) reported that all provinces conduct some form of assessment of student skills annually either using tests developed by themselves or published standardized tests. But as in the U.S., teacher-made tests form by far the bulk of testing that students encounter. As an example, Webster (1987) determined that in a large urban school division in Manitoba, "the majority of teachers prefer observation and teacher-made tests as assessment techniques" (p. 67), suggesting that teacher-initiated assessment forms the bulk of their assessment activity. In fact, assessment by observation (anecdotal) was used at least once a month by 36% of teachers and once a week by 33%; teacher-made tests were used at least once a month by 54% and once a week by 28% (Webster, 1987). In British Columbia, J. O. Anderson (1989) concluded "that most tests a student encounters are developed by classroom teachers to assess student achievement in specific instructional programs" (p. 125).

It is clear that classroom assessment is dominated by teacher-developed tests and other forms of assessment materials. Commercial standardized tests and other externally developed and mandated assessments are important in the overall evaluation that goes on in schools, but to a much lesser extent. The review that follows begins with a brief comment on the background to research in classroom assessment and a discussion of the role of standardized tests in classrooms. This is followed by a more extensive review of:

Teachers' classroom assessment practices
    In-class assessment time
    Classroom assessment purposes and practices
    Quality of teacher-made assessments
Teacher training in measurement.

The final sections of the chapter includes a summary, and the purposes and focus of the study.

## Background to Classroom Assessment Practices

Until recently very little was known about how evaluation was carried out in the classroom. As late as 1980, Lazar-Morrison et al. concluded from a review of the available test use literature that "while these reports indicate there is information about standardized achievement tests, little is known about the amount of other testing that takes place" (p. 6). Rudman, Kelly, Wanous, Mehrens, Clark, and Porter (1980), after reviewing literature from 1922 to 1980, indicated that prescriptive information on testing practices abounds but descriptive data on in-class testing practices is lacking. McKee and Manning-Curtis (1982) also commented that despite the importance of teacher-constructed tests "we know very little about the procedures instructors use to develop their tests" yet "much of course has been written about the *correct* way to plan and write tests" (p. 6). This paucity of information on classroom testing practices was echoed by Herman and Dorr-Bremme (1983), who attempted to determine the use of test results for decision-making by teachers and principals. It is apparent that even in recent years much of the research still tended to focus on the use of standardized tests--so much so that Stiggins, Conklin, and Bridgeford (1986) stated:

> Nearly all major studies of testing in the schools have focused on the role of standardized tests. . . . A review of the four most recent volumes of the [*Journal of Educational Measurement*] conducted by Stiggins and Bridgeford (1985) revealed that nearly all reports on achievement measurement dealt with topics relevant to the use of paper and pencil tests, and the vast majority of those focused on topics most relevant to large-scale standardized tests. (pp. 5, 6)

Much has changed in the last few years, and a substantial body of literature is accumulating on a number of topics relating to assessment in schools. The research agenda is broadening to include work on classroom-oriented evaluation (e.g., J. O. Anderson, 1990; Stiggins, 1990). But it is necessary, nevertheless, to review the role of standardized forms of assessment in classrooms.

### Teachers and Standardized Test Use

The earlier literature on testing indicated that very little use was made in the classroom of standardized achievement test results for any of the various possible purposes. In her 1971 review Kirkland concluded that many of the alleged social and psychological effects of standardized tests on students and schools were not founded on empirical research: much had been written but there were few reports on empirical studies of actual impacts of testing or even of the experiences or opinions of educators and others. What she did find was that principals were generally supportive of standardized testing and from 37% to 50% of them found the results useful for comparing student performance in their schools' performance with that of students other schools, and to interpret this to the public (Kirkland, 1971). Goslin (1967) concluded from a study of elementary and secondary teachers that standardized test results were used primarily to diagnose individual difficulties and to provide information to the student, but reliance on these types of data was minimal: less than 20% of teachers had altered a course and less than one third reported changing their methods as a result of tests.

Standardized testing in schools has received considerable criticism from various groups, notably the National Education Association's 1972 call for a moratorium on administration of certain standardized tests in schools (see Houts, 1977). Stetz and Beck (1979) rationalized their large survey of teachers' and students' attitudes toward standardized achievement testing on the basis that research to date was primarily on ability testing and included mostly persons outside of the classroom: parents, counsellors, and principals. Overall they determined that more than 90% of teachers thought standardized achievement tests were to some extent helpful and useful, although this included approximately two thirds of the teachers who responded 3, 4, or 5 on the 7-point scales used to obtain the teacher ratings: not powerful support for standardized testing. However, fewer than 20% of teachers favoured the moratorium on standardized achievement tests, 26% on intelligence tests, and 31% on state-mandated achievement tests. Stetz and Beck (1981) further reported that only 10% of the teachers surveyed made no personal use of standardized achievement test results, whereas 9% rated their use as "considerable", 50% rated it "some", and 31% rated it "little". This use was for a variety of purposes (see Table 1 below), although actual amount of use cannot be determined from the data. They concluded that teachers are not as negative about the use of standardized tests in schools as the critics of standardized testing might claim, or as measurement specialists believe that teachers are.

**Table 1.** Percentage of Teachers Reporting Use of Standardized Achievement Test Results in Their Classrooms[a]

| Type of classroom use | Percent reporting use | Type of classroom use | Percent reporting use |
|---|---|---|---|
| individual student evaluation | 65% | evaluation of teaching methods | 37% |
| diagnosing strengths and weaknesses | 74 | reporting to parents | 42 |
| class evaluation | 45 | reporting to students | 24 |
| instructional planning | 52 | measuring "growth" | 66 |

[a]Abstracted from Stetz and Beck (1981, p. 6).

Based on their review of the Royal Oak Study in Michigan completed by Boyd, Jacobsen, McKenna, Stake, and Yachinsky (1975), Lazar-Morrison et al. (1980) concluded that "although teachers reported variable use of results from the district-mandated testing program, there was little evidence that the testing program influenced curriculum or classroom instruction. . . . teachers felt that normed, standardized achievement tests were selected by administrators and imposed on teachers, and did not furnish them with any new information" (pp. 5, 6). Shulman (1980) concluded that of ten teachers who had received detailed diagnostic test data on individual students "not a single one of the ten teachers had looked at these results. They simply did not find them useful. . . . Most of the teachers did not believe the tests were of any value" (p. 68). And these were state-mandated domain-referenced diagnostic testing results for individual students!

Sproull and Zubrow (1981) spoke with system administrators and reported that at the district level, of the three main categories of standardized tests, achievement and ability tests were used with approximately equal frequency, and vocational aptitude tests with much lower frequency. Central office administrators thought that standardized test scores were used most frequently for individual student diagnosis and placement, instructional program evaluation, end-of-year achievement measurement, and reporting to outside audiences, with the student-oriented purposes receiving the greater emphasis. They attributed the primary benefits of test use to others, to principals and teachers, and apparently make only casual use of standardized test results themselves (e.g., comparisons of performance with that of previous years). Standardized test scores clearly did not dominate the information base for their decision-making. These authors further reported that two common purposes of the district-wide standardized tests were individual student diagnosis/placement and reporting to outside audiences, and that some internal use of the test data is made by district-level educators either through comparisons of present district performance with that of previous years or comparisons with national norms.

Salmon-Cox (1981), who interviewed elementary school teachers in Pittsburgh, found "that teachers, when talking of how they assess their students, most frequently mention 'observation' . . . [and] also mentioned teacher-made tests and 'interaction' with students . . . yet only three of 68 teachers mentioned standardized tests" (p. 632). It is informative to note that only 21 of these teachers felt they would care if standardized tests were abolished, and 8 more felt other teachers would care but not they themselves. However, teachers assumed that others outside the classroom would miss the test scores, others such as parents and school administrators. These results, although consistent with the Lazar-Morrison et al. (1980) and Shulman (1980) conclusions above, form a disconcerting paradox when coupled with the views of central office administrators presented by Sproull and Zubrow (1981). The findings are also in contrast with Stetz and Beck's (1981) more sanguine portrayal of teachers' views of standardized tests.

The use of standardized test results, as reported by teachers in the Salmon-Cox (1981) sample, was not only limited but was primarily as a supplement to or confirmation of a teacher's personal evaluation of the student. The teachers indicated little instructional use of these types of tests (e.g., for reshaping instructional content) or for grouping students for instruction (again this appears in sharp contrast with the findings of Stetz & Beck, 1981). Finally, the teachers tended to feel more confident with their own judgments of children's classroom performance than those derived from standardized achievement test scores: the test scores tended to be discounted if they were below what the teacher would predict, and to serve as a red flag to reconsider the teacher judgment if the test score was higher (Salmon-Cox, 1981). However, teachers did evince desire for diagnostic instruments, while they did not report standardized achievement tests as being useful for this purpose.

Madaus (1981) found that the opinions of teachers in Ireland were similar to those reported by Stetz and Beck (1979), and that standardized tests were not typically perceived in a negative light, at least initially. His results agreed with those of Sproull and Zubrow (1981) in that central office administrators and principals made little use of test results. But the experience of standardized testing and norm-referenced reporting appeared to diminish teachers' opinions as to the usefulness of this type of information. An earlier report by Airasian, Kellaghan, Madaus, and Pedulla (1977), based on the same Irish teacher sample, agreed with the Salmon-Cox (1981) findings that teachers make positive shifts of student ratings based on standardized test results rather than negative ones, although the number of changes was less than 10% .

Yeh, Herman, and Rudner (1981) stated that 58% of teachers reported "[standardized] test results were most important for initial reading placement, and 66% used test results most often for initial mathematics placement" (p. 6), but reliance on test scores reduced as the school year progressed--interactions with students and informal assessments (e.g., oral questions and quizzes) and teacher-developed assessments became heavily used. Standardized tests and curriculum embedded tests were little used. These findings were in keeping with those of Salmon-Cox (1981), and in contrast with those of Stetz and Beck (1981): standardized test use by teachers was primarily for reporting to parents or for evaluating materials, and even this use was very little; also, teachers felt positive to discontinuation of testing.

Fennessy (1982) reported that for elementary school teachers in the state of Victoria, Australia, by far the most frequent use for assessment was of non-structured observation, followed by teacher-made written tests, oral tests, and checklists. Diagnostic tests, standardized achievement tests, and screening tests were used much less frequently with 19%, 32%, and 32% of teachers, respectively, reporting never using these three types of tests, and a further 38%, 26%, and 31% respectively reporting rare use only (Fennessy, 1982). This is strikingly similar to the pattern of assessment practice reported by Salmon-Cox (1981) and by Yeh, Herman, and Rudner (1981).

Gullickson (1982, 1985) found that the role of standardized tests in teachers' evaluation of students in Grades 3, 7, and 10 was rated far lower by teachers than the role of teacher-made objective tests, and lower than essay tests and oral quizzes as well in Grades 7 and 10, although some of these differences were not statistically significant. He also found consistently lower mean ratings across grade levels for standardized tests than for teacher-made objective tests, essay tests, and oral quizzes for the content areas of science, social studies, and language arts.

Herman and Dorr-Bremme (1983) reported from a nationwide sample of elementary and secondary principals and teachers (English and mathematics teachers at the Grade 10 level) that both elementary and secondary principals rated the importance of standardized test results, including those of minimum competency testing and of district objective-based testing, clearly lower than teacher-made and curriculum tests or teacher opinions/recommendations for these decision areas: student class assignments, teacher evaluation, allocating funds, communicating to parents, and student promotion--primarily school- and class-level functions. The decision areas of public communication and reporting to district received higher ratings for standardized, norm-referenced tests than for teacher tests or teacher opinions/recommendations by elementary principals, and higher ratings for minimum competency tests by secondary principals. Teachers rated their teacher-made tests and observations/opinions considerably more important than standardized tests (including norm-referenced, minimum competency, and tests included with the curriculum) for four in-class decision areas: instructional planning, student grouping, remediation, and grading. It is apparent that standardized tests, including a variety of types, were considered important by both principals and teachers primarily for district and public reporting, as was found by Salmon-Cox (1981) and Yeh, Herman, and Rudner (1981), and clearly not for in-school or in-class decision making, in contrast to what was expected by district administrators according to Sproull and Zubrow (1981). These findings are also somewhat different from those of Stetz and Beck (1981).

More recently, Stiggins and Bridgeford (1985) reported that between 25% and 44% of teachers in grades 2, 5, 8 , and 11 indicated no use whatsoever of published tests (this included standardized objective achievement tests and objective tests supplied as part of published text materials--curriculum embedded tests). Their percentages of nonuse are similar to those reported for elementary teachers in Australia by Fennessy (1982). The substantial percentages of nonuse are higher than those reported by Stetz and Beck (1981), who reported 10%, and certainly reflect the low level of importance attributed to the use of standardized tests in the classroom reported by other researchers: Gullickson (1985); Herman and Dorr-Bremme (1983); Salmon-Cox (1981); Shulman (1980); and Yeh, Herman, and Rudner (1981). Level of nonuse varied from 15% in mathematics to 34% in science and in writing and 54% in speaking (Stiggins & Bridgeford, 1985). It is disconcerting to note that the relative importance teachers assigned to published tests was independent of the purposes identified for the testing: diagnosis, grouping for instruction, assigning grades, evaluating effectiveness of instruction, and reporting to parents. However, Stetz and Beck (1981) did find differences in proportions of teachers who reported using standardized tests for various purposes (see Table 1 above).

In the Monitoring Achievement in Pittsburgh program reported by LeMahieu (1984) and LeMahieu and Wallace (1986) the objectives and test items were developed with district support and with the input and assistance of the teachers involved, and, therefore, the tests ought to have been directly linked to the objectives (or perceived to be so by the teachers). Students were tested four to six times per year and schools were assisted in the interpretation of the results. Clear and substantial achievement gains on standardized norm-referenced tests were attributed to the program (LeMahieu, 1984; LeMahieu & Leinhardt, 1985). These gains may be attributable to the monitoring program. Cohen (1987) and Cohen and Hyman (1991) reported considerable gains in student achievement as a result of alignment between instruction and the test: "the more precise the alignment, the better the instruction and the higher the test score" (Cohen & Hyman, 1991, p. 20). The quality of the learning, as well as the ethics of teaching directly to the test, have been questioned as being too narrow a focus for instruction and as representing inappropriate preparation of students for the test (Mehrens & Kaminski, 1989). These tests were clearly high-stakes, and resulted in what is described as measurement driven instruction (Airasian, 1988b; Kirst, 1991a, b; Popham, 1987). What is unclear from these studies is exactly how teachers use the assessments in the classroom.

Hall, Carroll, and Comer (1988) reported that teachers emphasized teacher-prepared tests over externally-prepared tests for purposes such as providing evidence of student progress and assisting in promotion decisions. The emphasis did not appear to be great, with a mean of 3.7 on a 5-point scale in favour of teacher-prepared assessment versus 2.7 for state competency tests, for example. These results gave no clear indication of how much various test types would be taken into consideration in classroom decision making, particularly since other forms of assessment (marking of assignments, observation, etc.) were not included in the ratings.

In Canada, McLean (1985) concluded "careful, persistent questioning of officials and teachers at every opportunity revealed very few uses of this [norm-referenced] test information. . . . and in a strong majority of schools are never studied. Teachers do not regard them as relevant to their curriculum" (p. 37). This conclusion is supported by the findings of J. O. Anderson (1989) and Webster (1987), and in a general way by those of R. J. Wilson (1990), who reported that although teachers obtain assessment instruments from external sources, these are primarily workbooks and texts and the assessments often are modified by the teacher prior to use. It appears safe to conclude that very little if any use of standardized tests is made by teachers in Canada as well.

Burstall (1986) reported that teachers are supportive of the Applied Performance Assessment program in the United Kingdom. Teacher support is attributed, in part, to the nature of the assessment materials, which included practical components, administered on a one-to-one basis, for the areas of mathematics and science, and an oracy component in language arts. Support for the program may also have been gained from the fact that only group/school level information is reported, and the results are not for teacher evaluation. This puts into question how the tests and test results are actually used: what is done as a result of the information? Classroom-level decision making would be difficult if only school-level information is supplied. Earlier researchers had noted very little program or curriculum change by teachers on the basis of standardized test results reported back to the school (e.g., Boyd et al., 1975; Salmon-Cox, 1981).

## Comments on Standardized Test Use in the Classroom

Boyd et al. (1975) noted a number of problems relating to the use of externally developed tests by teachers: minimal teacher involvement in the testing program, lack of clarity and explanation of the purposes of the testing, lack of preparation in procedures for administering the tests and interpreting the results, lack of congruity between content of the test and what is being taught, and inordinate requirements of time for the testing. LeMahieu and Wallace (1986) argued that for the test information to be useful diagnostically it must be "relevant, appropriate, and timely" (p. 13). Three major hypotheses have been advanced as to why teachers have little interest in standardized test results: (a) there may be no incentive to use the results to change instructional practice (this was noted by Madaus, 1981, for administrators, but could also be true of teachers); (b) the tests do not provide teachers with information that they do not already possess, or perceive they possess (e.g., Madaus, 1981; Salmon-Cox, 1981); and (c) standardized tests do not provide teachers with the timely, teaching-sensitive information they appear to need for their in-class use (e.g., Salmon-Cox, 1981; Webster, 1987). If there is no incentive nor assistance in utilizing the results and in making changes, very little is likely to occur. Teachers reported having greater faith in their own assessments than in those based on standardized tests (e.g., Salmon-Cox, 1981; Stiggins & Bridgeford, 1985). They tended to see standardized tests as not being valid, either for decisions about the curriculum and its general implementation or for instructional decisions within the context of the teacher's unique interpretation of the curriculum and particular classroom and students. The tests simply do not fit exactly the instruction in a particular classroom at a particular time: "there is not likely to be a perfect match between the content of a standardized achievement test and a set of local curricular objectives" (Mehrens, 1984, p. 9). To obtain instructional effects curriculum alignment is a key feature, and this is most likely to occur if the teachers are involved in the test development process (LeMahieu, 1984; LeMahieu & Leinhardt, 1985). If we want teachers to use standardized assessments in the classroom then they should be involved in the development and construction of the tests.

State- or province-developed achievement tests, including minimum competency tests, have the potential to be more appropriate to classroom use (curricularly aligned) since they can accommodate regional variations in curriculum and instruction, but in the eyes of teachers they seem to fare little better than do commercial tests (e.g., J. O. Anderson, 1989; Haertel, 1986). They could have considerable potential for long range course planning and perhaps school program planning. They have strong political and public backing and this should be incentive for utilization (Kirst, 1991a). This should also be true of curriculum-embedded tests, tests that accompany curriculum and learning materials. However, as with norm-referenced achievement tests, they do not enjoy much enthusiasm from teachers. Stiggins and Bridgeford (1985) indicated very little use of standardized tests including test items accompanying text materials, and Yeh, Herman, and Rudner (1981) found little or no use reported by elementary teachers of standardized tests which included curriculum-embedded tests. Both elementary and secondary teachers in the Herman and Dorr-Bremme (1983) survey rated standardized tests, including minimum competency tests and district objective-based tests, substantially lower than teacher-made tests and teacher observations. Haertel (1986) concluded from his survey of secondary teachers that "only 32 percent thought that overall, district-wide standardized achievement tests were worth the time and effort they required to administer, score, and interpret. . . . 41 percent were unsure; 28% said that these tests were not worth the time and effort" (p. 3). J. O. Anderson (1987) reported that Grade 4, 7, and 10 science teachers in British Columbia indicated little or no emphasis placed on standardized tests for grading purposes--this included provincially developed achievement tests. It is reasonable to expect that teachers would not use commercial standardized tests as part of their student grading procedures, but this makes less sense at the grade 11 level since the British Columbia tests were provincial achievement tests designed specifically for the curriculum at particular high school courses and grade levels.

Part of the problem in determining classroom use of standardized tests is the lack of precision in the use of terms such as minimum competency tests and curriculum-embedded tests (e.g., David, 1979). The literature does not distinguish clearly among various types of either of these test categories, and teachers are generally asked to respond to a broad group of tests which includes tests that may or may not be properly subsumed under the headings. Teachers may also misunderstand what is meant by various categories of tests. Thus it is difficult to determine what use is made of externally-developed tests, and to what exactly teachers are responding in the various surveys that have been conducted: norm-referenced achievement tests, criterion-referenced tests of competencies, tests embedded in textual materials, a combination of these three types, or a combination of these and other types. This ambiguity becomes more problematic since there are reports of instructional impact as a result of a large-scale assessment program (Cohen, 1987; LeMahieu, 1984).

Although there is ambiguity in terminology it is apparent that there is a general lack of enthusiasm by teachers for all variations of standardized tests. In spite of this, the textbooks and professional journals still devote a surprising amount of space to standardized testing. This is changing, as Stiggins (1990) notes: "we have made a complete transition from caring totally about standardized tests to caring more about better classroom assessments and better classroom-level decision making" (p. 92).

## Teachers' Classroom Assessment Practices

The preeminence of teacher-developed assessment in the classroom has been well documented, particularly in the recent measurement literature. The importance of classroom assessment is discussed below in terms of the time and effort devoted to it. This is followed by a review of actual assessment purposes and practices.

### In-class Assessment Time

The importance of student assessment, for whatever purposes, is readily apparent from the substantial amount of classroom time devoted to testing. For example, Haertel (1986) reported that in high schools "it appears that on the average, at least 10 to 15 percent of class time available for instruction is devoted to paper-and-pencil testing, and . . . in some classes, well over 20 percent of class time is spent on testing" (p. 4). These estimates did not include the time spent on discussion of the results of the tests and quizzes, nor did it include the time devoted to evaluation based on various exercises and assignments, classroom behaviours, or informal questions--in short, it included only the more formal testing time. Surveys in the U.S. generally confirmed that from 10 to 20 percent of classroom time is taken up by testing (e.g., Carlberg, 1981; Gullickson, 1982; Newman & Stallings, 1982). Gullickson (1982) and Webster (1987) noted that probably even a higher percentage of total teacher work time is given to test preparation, construction, marking, and so on. These results indicate that from 10 to 27 hours are used each year for test administration alone (e.g., Dorr-Bremme & Herman, 1986).

There is considerable teacher and school variation in time devoted to assessment, but in general, testing time increases with grade level and appears to be greater in areas such as reading and mathematics than others (Dorr-Bremme & Herman, 1986). Some form of assessment is going on for more than the estimated testing time, and Green and Stager (1986) concluded that on average 40-50% of course grades are based on test scores (although across teachers the percentages range from 0% to 100%), the remaining percentage being based on other forms of assessment. They concluded that "time per week in testing-related activities" averaged 4.7, 7.4, and 6.9 hours for elementary, junior high, and high school teachers respectively. Stiggins (1987a) states

> teachers may spend as much as a quarter to a third of available instructional time directly involved in assessment related activities. . . . this includes time spent developing (or selecting), administering, scoring, recording and reporting daily assignments, paper and pencil tests and quizzes, text-embedded assessments, performance assessments (based on observation and judgment), and oral questions measuring achievement, ability and affective characteristics of students" (p. 3).

More recently, in Canada R. J. Wilson (1989) reported that "some evaluation would occur approximately every nine hours" (p. 137), and suggests that for most students this means an evaluation in one course or another would occur approximately every second day. In high school calculus, Traub, Nagy, MacRury, and Klaiman (1989) concluded that teachers spent on average 10% of classroom time on assessment (ranging from 8 to 16%), but that an average time of 11% was spent on review. Finally, although it was not the focus of their study, Boothroyd, McMorris, and Pruzek (1992) reported "that 7th and 8th grade mathematics and science teachers frequently test students with teacher-made tests, approximately one test every two weeks per class. Teachers were found to place more weight on students' scores on these tests when assigning end-of-course grades than on other forms of assessment" (p. 7).

It appears safe to conclude that as much as an equivalent of one day out of five of classroom instruction time, and perhaps more at the secondary school level, is given over to evaluation and evaluation-related activities on average. As an example of what this means in the life of students, J. O. Anderson (1989) indicated that students report receiving in the neighbourhood of 500 to 600 tests over their school years. Marso and Pigge (1992) put the figures at "between 400 and 1000 teacher-made tests before graduating from high school" (p. 3), although they estimate time spent on assessment to be in the range of 5% to 15% of classroom time. The distinction between assessment and instruction is arbitrary at best, and learning certainly does occur during time spent where the focus is on assessment, but this does give an indication of the central role of assessment in education.

## Classroom Assessment Purposes and Practices

Classroom assessment can be described in a number of ways and analyzed according to attributes such as the purposes of the assessment, the form of the stimulus used, the formality of the assessment procedures, and the type of response elicited. The language used to describe evaluation in the classroom setting is clarified first. This is followed by a review of teacher assessment practices, with a separate section dealing specifically with teacher grading practices.

Describing classroom assessment. Some of the assessment activities of teachers fit under the heading of testing, as this term is commonly used in education, whereas others would probably be called marking assignments, rating products and behaviours, or simply observing. After interviewing and observing classroom teachers in action Lortie (1975) notes:

> The monitoring techniques available to teachers are limited in number and precision; essentially, they must rely on various tests of students' knowledge and on observations of how students behave in the classroom. Tests include teacher-prepared examinations, verbal quizzes, student workbooks, and standardized tests. Observing student behavior includes judging student interest, watching work effort, checking compliance, and noting the degree of responsiveness to the teacher. (p. 138)

The *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers et al., 1990) use *assessment* for the process of gathering and quantifying information and making quality judgments to aid in decisions. This use is similar to that for the term evaluation, and some measurement authors use the two terms interchangeably (e.g., Mehrens & Lehmann, 1984). It is also the way the term is used in the *Principles for Fair Student Assessment Practices for Education in Canada* (1993). The sense of the term assessment is usually narrower than that of evaluation, and restricted more to systematic and quantitative procedures, whereas evaluation encompasses any process of forming value judgements regarding a person, behaviour, or object--or any educational phenomenon. Scriven (1977) and S. B. Anderson, Ball, Murphy, and Associates (1975) state that assessment typically has a quantitative orientation, and that it is often multidimensional and may incorporate a variety of test and nontest data. However, the term assessment is not used consistently by educators. Popham (1988) notes that some authors equate measurement with assessment, and in at least one measurement textbook assessment is treated as synonymous with testing and certainly not with evaluation (Wiersma & Jurs, 1985).

*Test* can be used in its broadest sense as systematically gathering and categorizing observations of a set of behaviours, and, typically, quantifying these categories according to a rule. This is behaviour measurement as described by authors such as Crocker and Algina (1986), Cronbach (1984), and Gronlund and Linn (1990). In classroom settings the term test is usually restricted to more formal assessment procedures for assessing achievement, often using paper-and-pencil tasks: this is the preferred use by writers such as Mehrens and Lehmann (1984) and Ebel and Frisbie (1986). Terms such as instrument, scale, and schedule are used for devices that assist in measuring student behaviours and products and student attributes of an affective or typical performance nature. Less formal assessment procedures may include interviews and observations (e.g., Gronlund & Linn, 1990; Mehrens & Lehmann, 1991). Usage of all these terms is quite variable from one textbook to another, though.

All the assessment procedures mentioned above were included under the rubric of psychological testing by Cronbach (1984) and Anastasi (1988), but the test-nontest assessment distinction is made in the educational literature, which more closely corresponds with teacher usage. Teacher-oriented language is probably more useful for discussing the assessment practices in classrooms: tests refer to the formal assessment approaches where students receive a standard or common set of stimuli and their responses to these stimuli are marked and the marks tallied according to some scheme. In the main, classroom tests are of the paper-and-pencil variety. Assessment includes tests but also the broader range of evaluation activities which involve quantification . Evaluation using subjective techniques, and evaluation of actual performances and affective-type behaviours are considered assessment as well (American Federation of Teachers et al., 1990).

Thus, *classroom assessment* is used here to encompass all the various forms and types of evaluation carried out in the classroom by teachers. The term *test* refers to more formal types of assessment, those conducted in a structured manner in the classroom, and usually involving paper-and-pencil tasks.

The purposes and uses of classroom assessment are also varied and diverse, and researchers describe them in differing ways. For example, Stetz and Beck (1981) used categories such as student evaluation, class evaluation, measuring growth, and so on, which are clearly not discrete. Herman and Dorr-Bremme (1983) described decision areas such as planning teaching and determining grades. Most educational measurement textbooks delineate a number of purposes which classroom assessments typically serve. For example, Nitko (1983) specified tests as being useful for these decision areas (which include grading and providing feedback): selection, placement, classification, counseling and guidance, educational diagnostics and remediation, and program improvement and evaluation. Gronlund and Linn (1990) provided a similar list of test uses, but identify four major categories of evaluation purposes: placement, formative, diagnostic, and summative. These categorizations, although useful generally, are not sufficiently detailed to describe teacher practices.

Stiggins (1987a) provided a more extensive list and outlined the questions that can be posed to determine teacher practice with respect to each use. The list included the categories outlined by Nitko (1983) and Gronlund (1985), but gave greater detail and was expressed in categories which are probably more readily adaptable to teacher practices. This list and the accompanying questions were used as a guide for interviewing the teachers in the case studies (from Stiggins, 1987a, pp. 14-17):

A. Diagnosing individual needs of students
B. Diagnosing group needs
C. Assigning grades
D. Grouping for instruction within class
E. Identifying students for special services
F. Controlling and motivating students
G. Evaluating instruction
H. Communicating achievement expectations
I. Communicating affective or behavioural expectations
J. Providing test taking experience.

Several additional distinctions regarding the form and type of assessment activity are important to note. The traditional categorization of tests into objective types (e.g., short answer, multiple-choice, matching items) and essay types (essentially any long answer paper-and-pencil items) distinguishes test items for which scoring is independent of the scorer from those for which it is not. This distinction is not precise nor is it consistently applied in measurement textbooks. Most writers include short-answer and completion items in the objective category (e.g., Ahmann & Glock, 1981; Ebel & Frisbie, 1986; Mehrens & Lehmann, 1984; Gronlund & Linn, 1990; Bloom, Madaus, & Hastings, 1981) whereas some do not, and treat these item types as a separate category (e.g., Nitko, 1983; R. L. Thorndike & Hagen, 1977). Several authors noted that completion and short-answer item formats may be somewhat less objectively scored than the choice-type item formats yet kept them in the objective category (Ahmann & Glock, 1981; Mehrens & Lehmann, 1984).

Assessment items can be categorized according to the type of response required from the student: choice- or selection-type responses (student selects the correct or best response from several that are supplied, such as in a matching exercise), and constructed- or supply-type responses (student constructs the response, such as in an essay item). This categorization is used to describe item formats in a number of test construction texts (e.g., Mehrens & Lehmann, 1984, 1991; Nitko, 1983; Popham, 1990; R. L. Thorndike & Hagen, 1977). Both Gronlund and Linn (1990) and Ahmann and Glock (1981) distinguish between supply and selection types of objective items.

Stiggins and Bridgeford (1985) delineate procedures based on direct evaluation of student performances or products as applied performance testing. This type of assessment can include paper-and-pencil tasks such as essays or stories, but only if students produce an actual composition rather than something formatted and structured by the teacher. Applied performance testing includes assessment of such things as laboratory skills, behaviours, and reports; research projects; and composite physical skills. Applied performance testing appears to be highly valued by teachers (Stiggins, Conklin, & Bridgeford, 1986), yet there are only limited resources and guidelines available for these less well-defined techniques. Some authors have attempted to provide assistance to teachers in assessing various products and performances, and behaviours, in the classroom context: L. Anderson (1981) and Gable (1986) on attitude assessment; Berk (1986a), Priestley (1982), and Stiggins (1987b) on performance assessment; Cartwright and Cartwright (1985) and Haynes and Wilson (1979) on behaviour observation. Some commonly used measurement textbooks provide guidelines in these types of assessment procedures, but to varying degrees: notable are Gronlund and Linn (1990), Mehrens and Lehmann (1991), Nitko (1983), and Popham (1990). These alternative forms of assessment are receiving considerable interest today, partly because of the renewed emphasis on higher-level thinking skills (e.g., Nickerson, 1989a; Wolf, Bixby, Glenn, & Gardner, 1991; Wiggins, 1989b). Efforts are being made to develop procedures and provide guidelines for direct performance assessment (e.g., Arter, 1993; Baxter, Shavelson, Goldman, & Pine, 1992), and the measurement community is taking seriously the need to incorporate these procedures in large scale assessments (e.g., Annual meeting of the National Council on Measurement in Education, 1993).

It is useful also to distinguish those test items which are based on material presented in the test, irrespective of the format of the item itself, such as questions based on a piece of text, a diagram, a map, or a set of apparatus and materials. Items which are context-dependent in this way are important to the classroom since it is one technique teachers can use to develop items that assess higher-level thinking skills and process objectives (e.g., Haladyna, 1992). Items can be made dependent on interpretation of the information presented and not entirely on memorized information. Items of this nature form the basis of many reading tests, and skills in their preparation should be part of the repertoire of teachers (e.g., Carter, 1984, 1986; *Principles for Fair Student Assessment Practices for Education in Canada*, 1993). An example of this type of testing material in paper-and-pencil format is the objective interpretive exercise described by Gronlund and Linn (1990).

Traditional item formats are also important to note: short-answer, completion, true-false, multiple-choice, matching, and essay. Nitko (1983) provides a comprehensive list of formats, which includes the typical ones listed above as well as those which would be included in Stiggins' (1987b) applied performance testing. Below is a slight modification of that given in Nitko (1983, p. 131):

Paper-and-pencil assessment
- Choice formats (selection-type)
    - true-false
    - multiple-choice
    - matching exercises
    - other alternate-response items
        - (e.g., key response)
- Short answer/completion formats
- Essay formats
    - restricted response
    - extended response

Behaviour and product assessment
- Performance observation formats
    - checklists
    - rating scales
    - sign and category systems
- Interviews, in-depth discussions
- Long-term activity formats
    - projects
    - extended written assignments
    - laboratory exercises
- Behaviour observations/ratings, affective
    - anecdotal
    - structured

**Teacher assessment practices.** There is recent research which attempted to determine the purpose and use of teachers' assessments, and their frequency, type, and format . Teachers rarely make use of standardized commercial tests or published tests available from other sources; however, the following discussion makes comparisons of use of teacher-constructed tests with that of standardized tests where this is appropriate.

Lortie (1975) interviewed 94 teachers from all school levels and reported that 47% used both tests and observations to monitor their teaching, 18% used tests only, and 24% used observations only. He further noted that there is considerable variety in patterns of assessment use: from near daily quizzes and continual observations of work samples to more formal and less frequent tests. In 1978 Yeh reported that of 260 K-6 teachers surveyed, 55% constructed their own tests, and that teachers tended to rely more on informal mechanisms such as observation and interactions with students for assessing student progress. She reported that teachers gave more tests in mathematics than in reading: the majority of teachers give mathematics tests weekly or daily and 80% at least monthly, whereas one third of teachers give reading tests weekly and another third do so monthly. Teachers in primary grades gave fewer tests than teachers in upper elementary. Yeh, Herman, and Rudner (1981) stated that teachers reported using their own tests to make instructional decisions, evaluate the classroom program, provide reporting information, and, somewhat less frequently, assign grades. The qualities described by teachers as important in external tests were: clear format, similarity to class material, and accurate prediction of achievement. Teachers cited suitability for their students and sensitivity to classroom instruction as major reasons for developing their own tests rather than using commercial ones (Yeh, Herman, & Rudner, 1981). Practical reasons such as costs and availability of materials also influenced teachers decisions to use their own tests.

Yeh, Herman, and Rudner (1981) reported that less experienced teachers were less likely to use external tests, of any kind, and relied more heavily on their own tests and observations. Teachers with assistants available to them reported greater use of curriculum-embedded tests and other techniques to monitor student progress. This may be due to the large amount of record keeping necessary for this type of monitoring which is facilitated by the classroom assistant. Also, it may be that testing is something which assistants can be assigned to do fairly readily, freeing the teachers to continue with teaching, as they prefer to do it.

Salmon-Cox (1981) found that teachers gave considerable emphasis to the development of social skills as well as to cognitive and affective learning. It was not clear how teachers determined progress in students' social skills, although observation was mentioned most frequently when talking about student assessment, and usually the typical objective format test could not be used for this purpose. Teacher observations, teacher-made tests, and interaction with students (oral questions) dominated classroom assessment practices. Salmon-Cox (1981) reported that teachers "get a feel" for their students fairly quickly: for example, 66% of them reported a general feel for students' abilities "within the first week or two or within the first month to nine weeks" (p. 632). The ability of teachers to judge students' abilities appears to be quite good when these judgements are made as ratings of students and compared to scores on standardized achievement and ability tests. Madaus (1981; Kellaghan, Madaus, & Airasian, 1980) found the agreement between students' scores on a standardized test (converted to a 5-point scale) and teacher ratings (on a 5-point scale) of students on the test constructs were very high--89% were within one point. More recently, Hoge and Coladarci (1989) concluded from a review of the literature that teachers' judgments of student achievement generally agree quite well with standardized achievement test scores (moderate to high correlations: medians in the .65 range), although variation among studies and among teachers was considerable. Mulholland and Berliner (1992) confirmed these conclusions, and further found that although experienced teachers were more accurate in predicting ITBS scores of their students, novice teachers could also be "remarkably accurate considering the fact that before the ITBS was administered, most of the novice teachers had between 16 to 20 hours of experience with the students they were judging" (p. 17).

Lazar-Morrison et al. (1980) concluded that "there is a great need to look at various types of assessments to determine the purpose they serve school personnel. Tests are apparently used for diagnostic, placement, grouping, and evaluation purposes but the specific tests used for these purposes is not known" (p. 9). These authors identified several factors related to the use of tests. The first was that teacher training in educational measurement and testing seemed to enhance positive attitudes to testing and increased the use of tests and test scores, although the effects of training were not clear (the research focus prior to 1980 was primarily on standardized testing). Based on (Yeh (1978), they further noted that more experienced teachers were positive toward tests and tended to use them more often (Yeh, 1978, was the basis for this point), and that situational factors may have had some bearing on testing practices. For example, the availability of classroom supports such as assistants seemed to increase the use of district-developed tests.

Lazar-Morrison et al. (1980) concluded teacher attitudes to testing seemed to have some bearing on practice. The Stetz and Beck (1979, 1981) results described earlier suggested only mild support for standardized testing; in part this may be because of the teacher belief that these tests were not valid in the classroom context.

Finally, Lazar-Morrison et al. (1980) argued that the level of teacher preparation in measurement was not high. This was argued to be the case by many measurement specialists both historically (e.g., Ebel 1967; Roeder 1973) and more recently (e.g., McLean, 1985; O'Sullivan & Chalnick, 1991; Rogers, 1991; Stiggins 1988a).

McKee and Manning-Curtis (1982) reported that instructors in a postsecondary institute for the deaf, most of whom did not have a teacher training background, wrote approximately 82% of their own test items, and the bulk of the remaining items came from other instructors' tests. They reported that short answer or completion items were used most frequently, followed by multiple choice, matching, true-false, and essay items. At first glance, these results are striking in that the essay format is least used, and this is at the college level. However, interpretation is difficult since one essay item can be equivalent in marks to many of the other items, and a major paper may be worth as much as 100 or more test items. The relative weights assigned for determining student grades gave a better indication of the importance associated with various types of assessment, and although tests were considered most important, papers, projects, and other assignments were also considered important (from Table 4 of McKee & Manning-Curtis, 1982): paper-and-pencil tests (41%), papers and class projects (21%), class participation (7%), laboratory tests and work (15%), and other (mostly homework) (12%).

Compared to a sample of high school teachers, the instructors were higher in their reported use of test construction principles: preparing a table of specifications, preparing the test one week in advance, making test length appropriate for time allowed, asking colleagues to review the test, calculating class mean, and calculating item difficulties (McKee & Manning-Curtis, 1982). Also, many of the instructors (40-50%) indicated an interest in inservice training in test construction and measurement techniques.

In Australia, Fennessy (1982) surveyed 116 elementary school teachers (65% return rate) and determined that nonstructured observations and teacher-made tests were the most commonly used assessment techniques (see Table 2 below). He found no relationship between previous measurement training and frequency of test type use, but obtained considerable differences in frequency of assessment across subject areas, the greatest frequency being in mathematics and reading, followed by written expression. Assessment in science, art and craft, music, health, and physical education was considerably less frequent (he considered the relatively low frequency of assessment in science as probably due to the fact that science is infrequently taught). These results are based on fairly young teachers, 70% under 30 years, and only 30% of them had a course in educational measurement.

**Table 2.** Percentages of Teachers Reporting Use of Various Assessment Techniques[a]

| Assessment technique used in the classroom | Never or rarely use | Often or very frequently use |
|---|---|---|
| nonstructured observation | 11% | 74% |
| teacher-made written tests | 8 | 73 |
| oral tests | 13 | 60 |
| checklists | 12 | 56 |
| diagnostic tests | 57 | 13 |
| standardized achievement tests | 58 | 7 |
| screening tests | 63 | 13 |

[a]Adapted from Fennessy (1982, p. 6).

Gullickson (1982) surveyed 336 South Dakota teachers (75% return rate) of Grades 3, 7, and 10 regarding their assessment practices in three curriculum areas: science, social science, and language arts. Only 5% of the teachers had no course(s) in educational measurement, approximately 57% had one course, and the remaining had two or more courses; also, 84% reported some measurement in other courses. Testing was a substantial part of their teaching activities with 95% of teachers reporting testing on a weekly basis and 98% on at least a biweekly basis. These percentages are considerably higher than the 65% estimated by Lortie (1975). Gullickson (1982) reported that teacher-made objective tests were the most important by far for assessment, then essay tests, followed by standardized objective tests and quizzes. Major roles were assigned to instructional feedback, evaluation of instruction, and grading. Motivating students, assessing attitudes, and providing for student input were assigned lesser roles.

Gullickson (1982) found 93% of teachers constructed their own test items, and 75% of Grade 3 teachers, 61% of Grade 7, 47% of Grade 10 reported using items supplied with textual materials. Only 23% reported using published test items and 12% using items prepared by other teachers. Teachers indicated using four to five different item types, with short answer/completion items being the most common, followed by matching and multiple-choice, and then by true-false and essay items. Teachers' ratings of importance of seven techniques as to their role in the evaluation of students were further analyzed and reported by Gullickson (1985). Objective teacher-made tests were assigned the greatest role by teachers at all three grade levels, and mean ratings were essentially equivalent for the three curriculum areas of science, social science, and language arts. The role of essays increased, and of oral quizzes and objective standardized tests decreased, from Grade 3 to 7 to 10. The differences for these techniques were small across curriculum areas, although essay tests had a slightly lower role in science than in the other two areas. The roles of papers or notebooks and discussion were the highest of the "nontest" techniques for all three grades and all three curriculum areas, although both were lowest in Grade 10 and in science. Laboratory ratings were, as can be expected, much higher in science than in social science or language arts. Projects and oral report ratings decreased slightly from Grade 3 to 7 to 10 and from language arts to social science to science. Finally, the role of citizenship in school was rated much higher than citizenship in community for all grades and curriculum areas, although the ratings of both decreased from Grade 3 to 7 to 10 and increased from science to social studies to language arts.

Interestingly, teachers in the Gullickson (1982) sample perceived tests to cover about 75% of the material they teach. Most of the testing was formal in nature: very few teachers allowed interaction during testing and there were few open-book tests. Finally, surprisingly, some 36% of the teachers required use of a separate answer sheet (although the amount of this use was not determined). This provides some idea of what teachers do in their assessments, but as commented regarding earlier research, it does not clearly specify the amount of various types and formats of testing, and gives little information on how and why specific assessment procedures are conducted.

Gullickson (1982) reported that teachers (78%) typically assigned a letter grade rather than just a numerical score to test results, and this grade was usually tied to a fixed proportion correct, such as 80-90% is an "A", and so on. Most teachers (90%) provided written feedback on the tests at least occasionally, with 55% reporting "always" or "usually". Test analysis, however, was typically not used: 42% of the teachers said they report the test score range, and only 10-13% the mean/median and standard deviation. A more recent analysis of these data by Gullickson and Ellwein (1985) led the authors to estimate that only 12% of teachers calculated and used reliability information, and 31% used item difficulty. Gullickson (1982) noted that most teachers returned test results to students rapidly (83% within one day). Teachers typically conducted test review following administration, about 20 minutes on average: 43% of this time was on the review of items identified by the teacher presumably from the item results, 41% was on item review based on student requests, and the remaining proportion of time was on grading practices. Elementary teachers "usually" allowed students to keep the returned tests and secondary teachers "sometimes" did so. Gullickson (1982) concluded that teacher assessment practices do not conform to practices as recommended by educational measurement texts: (a) the item formats most commonly used by teachers, short answer and matching, tend to test only lower cognitive levels of thinking; (b) few teachers systematically analyze and revise test items before reuse; and (c) teachers misuse the notion of criterion-referencing by not specifying clear domains and not establishing justifiable criterion points.

Chambers (1982) reported the results of an analysis of 23 randomly selected teacher-made tests in junior high social studies. On average there were 35 items per test, 38% of which were matching items, 22% completion or short-answer, 19% true-false, 15% multiple-choice, 3% essay, and 3% other types. Unfortunately, many of the tests did not show the point-value of each item; hence Chambers could not determine the relative proportion of assessment obtained by the various techniques.

Herman and Dorr-Bremme (1983) surveyed principals and teachers (return rate of 60% at the elementary level and 48% at the high school level) across the U.S. on assessment practices. They found that teachers rated teacher observations and opinions together as most important, with teacher-made tests a close second, over standardized test, district continuum or competency tests, and tests included with the curriculum for all purposes surveyed. Principals rated standardized tests as most important for areas such as public communication and reporting to districts. Both teachers and principals afforded standardized tests and other external tests some importance, but, lower than that for teacher tests and observations. Many teachers viewed tests as motivating students to study.

Stiggins and Bridgeford (1985) surveyed teachers from districts throughout the U.S. (60% response rate) on their use of various assessment techniques and the purposes of them. The major categories of techniques were specified as: teacher-made paper-and-pencil objective tests, published tests, structured performance assessment, and spontaneous performance assessment. Teachers reported making greater use of teacher-made objective tests at the higher grade levels and less use of published tests. They also reported making greater use of objective tests in science and mathematics than in writing and reading. This is consistent with what J. O. Anderson (1987, 1989) and Bateson (1990) reported: British Columbia science teachers at Grades 4, 7, and 10 all rated teacher-made objective tests as receiving the greatest emphasis for student grading purposes, and more so at the Grade 10 level. It appears from Stiggins and Bridgeford (1985), J. O. Anderson (1987, 1989), and Bateson (1990) that teachers are quite comfortable with both structured and spontaneous observation performance assessment. The pattern of both objective test and performance assessment use was similar across grade levels and subject areas with some increase in use of more formal testing in the higher grade levels (see also Stiggins, Griswold, & Wikelund, 1989).

Barnes (1985) determined from her observations and interviews that student teachers viewed evaluation as serving several purposes: motivating students, communicating to parents, classifying students, and assessing instructional effectiveness, whereas cooperating teachers (of the student teachers) tended to view the main purpose of evaluation as being a source for grading and reporting. She also was concerned that practicing teachers tended to include both effort and performance in awarding student grades. J. O. Anderson (1987, 1989), too, noted that some teachers include student effort and attitude in their evaluation for grades, although a substantial number indicated that they did not.

Webster (1987) surveyed and interviewed teachers from all grade levels and found that they use oral interview and work sample more frequently in their classroom assessments than they do teacher made tests or observation. She concluded that the assessment devices more commonly used by teachers are ones that are less time-consuming and that do not interfere with classroom activities as much as other devices might, and that this was a cost/time saving action on the part of teachers. She also concluded, in keeping with the results of Stiggins and Bridgeford (1985), that there was limited variation in teachers' assessment practices across grade levels and subject areas. Providing instructional and individual feedback, and assessing achievement were rated by the teachers as the most important reasons for classroom assessments (Webster, 1987). Also, teachers indicated that classroom assessments were important for understanding the child, both cognitive and affective, and for instructional modification.

Several recent studies in Canada provide further evidence of teachers' assessment practices. R. J. Wilson (1989) reported that in high school "evaluation instruments invariably served at least two purposes, and frequently more than two" (p. 140). The focus of classroom assessment was on making decisions about student progress, but the assessment results were often also used in a general way by teachers to review their courses and their instruction. Subsequently, he reported that teachers indicated the purpose of approximately 80% of their classroom assessment instruments as being for generating marks (R. J. Wilson, 1990). The purposes of practicing application of learning and checking students' progress were frequently endorsed by both elementary and secondary teachers, but diagnosis was primarily endorsed by elementary teachers. The frequency of assessments was fairly consistent across grade levels (K-13), but there was a general trend of moving from more performance assessments at the lower grade levels to paper-and-pencil techniques at the higher levels. Completion and short-answer items were by far the favoured format of the assessment instruments at all levels, but other forms like multiple choice, matching, essay, and performance appraisal were also common. The reports on the 1986 British Columbia provincial survey by J. O. Anderson (1987, 1989) and Bateson (1990) concured with these findings, but also indicated the importance of laboratory write-ups, research reports, and projects in the evaluation practices of science teachers.

Stiggins, Griswold, and Wikelund (1989) collected assessment material from U.S. teachers of all grade levels. Of the 4,120 assessment exercises "34% were selection-type, 54% were fill-ins, 10% were essay, and 2% required some other type of product or response" (Stiggins, Griswold, & Wikelund, 1989, p. 237). These results are generally in keeping with those of other researchers, but should be treated cautiously since only paper-and-pencil assessments were sought. In their extensive review of research into teachers and assessment, Marso and Pigge (1992) concluded that there are grade- and subject area-related differences in practices. The assessment becomes more formal at the higher grades, with more emphasis on testing and less on observation and work samples. Mathematics and science teachers also tend to test more frequently than do teachers of other subjects, and teachers of writing and speech rely more on direct observation and informal judgment.

Teacher grading practices. Until recently very little was known about how teachers determine grades for students (e.g., Manke & Loyd, 1990; Stiggins, Frisbie, & Griswold, 1989). However, the importance of grading to the lives of students is well recognized. The topic has received considerable attention from the measurement textbook writers, who give detailed accounts of the do's and don't's primarily based on principles associated with accurate and interpretable measurement (e.g., Ebel & Frisbie, 1991; Gronlund & Linn, 1990; Hills, 1981; Mehrens & Lehmann, 1991; Nitko, 1983; Popham, 1990). Manuals to guide teachers in their procedures for grading have been produced by Taylor (1979), and more recently by Stiggins (1991b) and Frisbie and Waltman (1992). These guidelines provide teachers with general principles of forming summative statements, but also outline the philosophical and practical considerations in deciding what form grades should take and how these might be constituted.

In 1977 the *Phi Delta Kappan* compared the views on grading practices of student teachers, *Kappan* readers, and National Council on Measurement in Education (NCME) members (see Nitko, 1983). There was some agreement among these groups on a number of issues, and decided differences as well. Over 50% of each of the three groups felt that student marks should be treated as measurements and not evaluations. Similar percentages in the three groups, approximately 35%, thought absolute standards are preferable to relative ones. However, student teachers (70%), and *Kappan* readers to a lesser extent, were more in agreement with variability in standards among teachers and over classes than were NCME members. More student teachers and NCME members thought achievement should be the basis of marks and not attitude and effort.

Stiggins, Frisbie, and Griswold (1989) and Terwilliger (1989) reaffirmed measurement specialists' views on grading practices, but recognized that teachers' expectations and judgements must be based on classroom experience and that perhaps what measurement specialists recommend may need to be tempered with other important considerations in reporting on student progress. Stiggins, Frisbie, and Griswold (1989) outlined some 19 grading practices recommended by measurement experts, and, based on case studies of 15 high school teachers, they found reported practices were inconsistent with 11 of these. Notable consistent practices were such things as making explicit the basis and procedures for grading; striving not to include student personality characteristics, and student attitudes toward and interest in the subject matter; using a variety of assessment procedures, including written and performance assessments; and not grading on the curve. Practices considered inconsistent with measurement specialists' views were such things as including in the formation of grades factors like motivation, effort, and ability; and classroom assignments designed for learning, and a host of other low quality assessment data (e.g., casual and informal observations); and also aggregating assessment data and setting cutoff points inappropriately; and deciding borderline cases using subjective non-achievement data.

Traub, Nagy, MacRury, and Klaiman (1988) found that even within one calculus course teachers used many different grading schemes, and included anywhere from 3 to 13 tests to form composite scores for grading. Some teachers relied entirely on tests, others included a variety of assignments and other tasks; reliance on tests ranged from 30% to 100% of the course marks. The procedures teachers used to calculate composite scores varied considerably also, and component weights did not consistently reflect course content emphasis or length of time spent. A significant, but not surprising finding, was that scores on final examinations averaged about 12% lower and tended to discriminate more than did term scores, and also averaged some 6% lower than midterm grades.

Bateson (1990) reported that for teachers of Grade 4, 7, and 10 science in British Columbia, tests form an increasing basis for marks as grade level increases. Most teachers reported that they had clear expectations of proportions of students who would receive various grades in their classes, and Bateson (1990) described these expectations as very similar to the grade distribution which would arise from the provincial examination in Grade 12 English. He also noted that one third of Grade 10 teachers indicated that they use a preset distribution of marks, which he argued was incompatible with the fact that they had classroom populations which varied from year to year. R. J. Wilson (1990) also found that teachers had well defined expectations of student performances and grades, and used these to establish pass-fail cutoffs.

The results of a survey of secondary school teachers by Friedman and Manley (1991) largely corroborated the findings of Stiggins, Frisbie, and Griswold (1989). The teachers generally agreed that although achievement should be the primary basis for grades, factors such as effort and motivation should be included in grading and in deciding borderline cases, but not interest and personality. They agreed that various assessment techniques should be used, but also agreed that data from daily work assignments should be included. Concerns for the technical aspects related to reliability and validity were largely ignored by teachers.

Manke and Loyd (1990) found that primary school teachers preferred narrative reports and parent-teacher conferences over letter grades. The teachers also stressed the importance of carefully considering the individual child in all evaluative statements. Secondary teachers stressed the importance of fairness, and that this included consideration of individual needs. They noted the importance of clear grading policies and explicit procedures. They also showed that teachers' evaluations may be affected by personal characteristics of students, such as gender and ethnic origin, and they supported the concern of measurement specialists that achievement only be considered in grading. Loyd, Nava, and Hearn (1991) reported that secondary students agreed with teachers on what they thought should be included in grades, that classroom assignments, and effort and motivation should be considered by teachers. One of the greatest concerns of these researchers was regarding potential bias and unfairness that this practice may increase.

The complexities in aggregating assessment data from several sources to form composites for grading have been well documented and guidelines for teachers are offered both in measurement textbooks (e.g., Hills, 1981) and elsewhere (e.g., Oosterhof, 1987; Stiggins, 1991b; Thayer, 1991). The research suggests that teachers do not typically use these procedures nor even understand them (e.g., Manke & Loyd, 1990, 1991). There is no evidence, for example, that teachers either explicitly or tacitly accommodate for the effects of differences in variances of the component assessments included in forming a composite.

The importance of grading cannot be overrated. Grades are not simply value interpretations of psychological measurements with theoretical implications. They represent judgements of the performance and prowess of students, and have huge implications for their lives. Cronbach (1977) notes that grades are never simple descriptions of student achievement, ". . .no one regards them that way. Students treasure them or worry over them; parents beam or scold. Teachers use marks deliberately to goad, reward, and punish" (p. 680). Stiggins (1991b) agrees: ". . . grades are emotional. . . . It is never a dispassionate, scientific act" (p. 2). Teachers often acknowledge that grading and reporting is one of the more difficult and perplexing tasks that they face in their jobs (Barnes, 1985; Marso & Pigge, 1992), and Terwilliger (1989) states "assigning grades to students is undoubtedly one of the most distasteful aspects of teaching" (p. 15).

Recent work has attempted to identify what teachers are trying to communicate with grades and how they view them. Brookhart (1992) commented, "grades functioned as the coin of the realm" (p. 4), and concluded that teachers emphasize the utility of grades as rewards for effort and accomplishment. Teachers tend also to be concerned about the consequences of the reported grades, and how this might affect students' future behaviour and learning. But teachers vary in this: some teachers seem to operate on the principle that students get what they earn, whereas others wish to motivate students and consider a variety of affective factors in giving a grade, yet all are concerned with fairness and equity (Brookhart 1992, 1993).

Teachers' grading practices apparently do not conform closely with the recommendations of measurement experts. The teaching situation is complex, and rarely can evaluative statements be made that are totally unambiguous. There must be judicious application of measurement specialist recommendations to classroom assessment, and this must take cognizance of the role of the teacher as well as the ecological milieu of the classroom, as Griswold and Griswold (1992) and Brookhart (1993) note. As well, the expectations for information and understandings of parents must be considered in how student performance is communicated (e.g., Waltman & Frisbie, 1993; Shepard & Bliem, 1993). As an example, the views of the meanings of grades vary from student to teacher to parent (Pilcher-Carlton & Oosterhof, 1993). Despite the importance of grading and reporting student progress, and how difficult it is perceived to be by teachers, relatively little empirical research exists and most of this is at the high school level, and no comprehensive, empirically grounded theoretical rationale is available to guide and assist teachers in their work.

One very important aspect of the whole process is communicating student performance to students, parents, and others. We have some information on forming grades, and on interpreting them, but as an example of the knowledge gap, we know virtually nothing about the parent-teacher conference except that it was a common feature in schools, particularly at the elementary level. Kunder and Porwoll (1977) reported that in the United States over 70% of elementary schools used this procedure, whereas approximately 35% of junior high and 26% of senior high schools did so. This study is now over 15 years old, but it points to the pervasiveness, if not the importance, of the parent-teacher conference. More recent information confirms that some form of parent-teacher conference is even more common in today's schools (Robinson & Craver, 1989; Wood, Bennett, & Wood, 1990). In fact, in some jurisdictions, at the lower grade levels these conferences are becoming the main mode of communicating student progress to parents, and grades and other forms of summary reports have been replaced with narrative/descriptive reports (J. O. Anderson & Bachor, 1993; Friedman & Frisbie, 1993; Schulz, 1993). We only have the beginnings of a research base to help inform and develop our advice for teachers, yet Waltman and Frisbie (1993) report parents and teachers "tended to agree that the regularly-scheduled parent-teacher conferences provided the single most useful information source" (p. 16).

## Quality of Teacher-Made Assessments

There are problems with the quality of teacher-made tests, and as these form the major part of evaluation in the schools this is cause for concern. The problems tend to be of two major types: the assessments do not assess many of the intended learning outcomes of a curriculum, especially those of a higher-level cognitive nature, and the assessments are often technically weak and may be unclear or may misdirect students. Curricula clearly call for learning of higher-level cognitive skills, and writers have identified approaches to their assessment (examples in science are Klopfer, 1981, and Yager, 1989; examples in critical thinking include Norris, 1989, and Quellmalz, 1985). Complete issues and major sections of journals have been devoted to the topic of assessing higher-level thinking and reasoning skills: for example, *Educational Researcher* (Nickerson, 1989a), *Educational Leadership* (Brandt, 1985, 1989), and *Phi Delta Kappan* (e.g., Haney & Madaus, 1989).

Fleming and Chambers (1983) reported from their detailed review of teacher-made tests in Cleveland, that, based on their judgements and those of trained teachers, the bulk of items were at the knowledge level of Bloom's cognitive domain (summarized in Bloom, Hastings, & Madaus, 1971): 67%, 94%, and 69% respectively at the elementary, junior high, and senior high levels. Furthermore, various administrative and technical aspects of tests, such as clear directions to students and relative weightings of items, were conspicuously absent from many of the test samples.

Carter (1984) found that secondary teachers could not readily recognize test items as requiring inferential reading skills. Only 29-33% of teachers correctly labeled two inferential reading items. They fared a little better with items requiring reading for detail: 50-68% correctly labeled these items. She also found that categorization agreement among teachers was poorer for items of higher order skills (e.g., 27% agreement for recognizing appropriate "inference" items) than for lower order ones (e.g., 55% for recognizing appropriate "detail" items). The teachers also took substantially longer to write "inference" items, approximately 24 minutes on average, than to write "detail" items, approximately 7.5 minutes on average (Carter, 1984, p. 58). Further interviews of these teachers revealed that they were insecure in their skills for producing tests of higher-level cognitive abilities, and Carter (1984) concluded that teacher training is wholly inadequate in preparing teachers for this task. In a later paper she showed that these teachers were unaware of faults in multiple-choice items, although Grade 7 students could capitalize on these faults to achieve well above "chance" without reading the passage on which the items were based--for example, 78% answered one item correctly rather than the expected 25% (Carter, 1986). Unfortunately, Carter did not make clear in either paper whether these teachers had any previous training or course work in educational measurement.

Haertel (1986) notes from his interviews of 15 high school teachers and reviews of their tests that the tests often required little more than repetition of textbook or classroom presented material or solution of problems exactly like those in class. He concluded that the teacher tests did appear to match the instructional content, but not necessarily the goals of the curriculum. General thinking objectives and affective outcomes were not assessed, nor did they appear in the objectives of those teachers who listed objectives. Typically "students were asked to apply their knowledge or understanding, and even items that appeared to call for analysis or supported argumentation proved in fact to require no more than reproduction of what had been said in class" (Haertel, 1986, p. 16).

Stiggins (1986a) determined that one teacher's interactive questioning during instruction tended to reflect all levels of Bloom's cognitive taxonomy but the actual tests administered to the children measured only simple recall (the tests used were provided as part of the textbook materials supplied by the text authors). Stiggins, Griswold, and Wikelund (1989) observed teachers from all grade levels and coded their oral questions according to the cognitive level that they were assumed to elicit (as these are defined by Stiggins, Rubel, & Quellmalz, 1988). The results showed an overall high proportion of questions at the recall level, ranging from 36% to 51% of questions for Grades 3 to 12 with no grade-related pattern. Grades 1-2 had 70% of questions at this level.

Stiggins, Griswold, and Wikelund (1989) also collected paper-and-pencil assessment material from the teachers and determined that the majority of items were at the recall level ( which includes Bloom's knowledge and comprehension levels). As with the oral questioning, this result was fairly uniform across grade levels with percentages of items ranging from 41% to 56%. There were substantial numbers of items rated at the inference level, primarily because many of the mathematics items, 72%, were judged to be in this category. However, when mathematics items were removed, the results for all grade levels were 55% recall, 16% analysis, 5% comparison, 19% inference, and 4% evaluation.

From their extensive analysis of over 6500 questions on tests produced by teachers who had taken a course in measurement, Marso and Pigge (1992) drew a number of conclusions.

1. Item type varied with grade level: completion and multiple-choice were more common at the lower grades; essays were very infrequently used at any level; and the most common formats generally were short answer, matching, multiple choice, and true-false.

2. Item quality was not related to years of teaching experience.

3. Matching exercises were the most error-prone.

4. When all the questions were combined, 72% were estimated to be at the knowledge level of Bloom's taxonomy, and the majority of those beyond the knowledge level were in mathematics and science (much as Stiggins, Griswold, & Wikelund, 1989, found).

It seems correct to conclude that the most common item types that teachers use on their tests are completion, short-answer, and simple selection-type items (e.g., Chambers, 1982; Marso & Pigge, 1992; Stiggins, Griswold, & Wikelund, 1989; R. J. Wilson, 1990). It should be noted that a recent report by McMorris and Boothroyd (1992) found science teachers to use a substantial number of multiple-choice items, although this was based on a rather small sample of test documents. Gullickson (1982) claims that completion, short-answer, matching, and true-false items are not the best to measure higher-level thinking. Although authors of measurement textbooks show how a variety of item types can be used to measure various learning outcomes (e.g., Gronlund & Linn, 1990), teachers apparently do not incorporate these in practice. This may be small wonder since authors of test materials to accompany textbooks appear to have difficulty producing higher-level thinking questions (Ellsworth, Dunnell, & Duell, 1990).

The lack of use by teachers of statistical techniques for analyzing their tests, as reported by Gullickson and Ellwein (1985), is but a small part of the problem. Item analysis statistics are useful to improve items, and may even help detect item biases, but cannot help in the initial drafting of good items that appropriately cover the content and domains of importance.

The studies reviewed above represent the results of relatively small numbers of teachers. However, the consistency of findings that assessment at the lower cognitive levels predominates paints a dismal picture. Educators such as Airasian (1988a, b), Shepard (Kirst, 1991a), Smith (1991; also Smith & Rottenberg, 1991), and Wiggins (1989a, b) claim that the effects of external assessments are to distort education and emphasize rote forms of learning. From these results it appears that there is little in classrooms to counteract this. Students at the secondary level, and some at earlier levels as well, recognize the source of data for their grades, and this is primarily teacher-made tests (e.g., J. O. Anderson, 1989; Crooks, 1988; Haertel, 1986; Herman & Dorr-Bremme, 1983). The signal to students is that recall-level learning is what is important!

## Summary of Classroom Assessment Practices

It seems clear to conclude from the research on classroom assessment practices that:

1. Teachers assess students frequently, both in formal and informal ways, but variation among teachers is considerable. Assessment administration and assessment-focused activity can encompass as much as one-fifth to one-third of classroom instructional time. This is true for all levels of the school system, although more formal assessments, tests, tend to be used more frequently at the higher grades and in science and mathematics.

2. Teachers assess for a variety of purposes, the most important being evaluation of individual students for judging their progress. This evaluation includes the grading and reporting function, but comprises providing feedback on student progress to students and teachers, and to parents more generally. Teachers typically use assessment results from one instrument for a number of purposes, some of which may be incompatible (e.g., summative grading and providing individual diagnostic information). Teachers prefer information which is directly related to their instruction and that is immediately relevant.

3. Formal classroom assessment tends to focus on cognitive skills. Assessment of student knowledge forms the basis of grading and reporting decisions. However, classroom assignments, long-term projects, and subjective appraisals of effort are all typically included in summative grades. The assessments tend to be heavily based on observation and ongoing work samples in the primary grades, but move more to formal paper-and-pencil testing in the higher grades. Teachers clearly prefer assessment procedures and materials that are developed directly by themselves, or those modified from other teachers' or existing textual materials.

4. The assessment format most favoured by teachers is completion and short answer, particularly on tests. There is also considerable use made of selection-type formats, particularly true-false and matching. These formats lend themselves to assessing different levels of cognitive functioning, but items found on teacher-made tests tend to emphasize lower-level skills (primarily the knowledge and comprehension levels of Bloom's taxonomy). Teachers use essay-type assessments and long-term projects to attempt to measure higher-level thinking, but often it is not clear that higher-level skills are what is being assessed. Quality control procedures are rarely used in schools, and there is little if any postmortem done on tests. Teachers do not typically review one anothers' assessments.

5. Grading practices vary considerably from school to school, across subject areas, and throughout the grade levels. Clear, explicit policies and procedures for grading and reporting are not the norm in schools. Teachers appear to use faulty practices in combining assessment information to form grades. Information from a variety of assessments is usually incorporated, including work assigned for student learning, and this is often not treated in such a way as to maintain the importance attributed to each component. Student effort and other factors are considered in cases or borderline grades.

## Teacher Training in Educational Assessment

Twenty-five years ago Robert Ebel (1967) lamented teachers' lack of adequate preparation, and with some justification. The same battle cry was as clear two decades later (e.g., Gulliksen, 1986, b; Jett & Schafer, 1992; Nitko, 1991a; Stiggins, 1985, 1988a, 1991a). Goslin (1967) reported that, at the time, in the U.S. less than 40% of teachers had more than minimal exposure to formal training in tests and measurement. Roeder (1973) reported that of over 800 institutions in the U.S. which indicated preparing elementary school teachers, only 31% required students to take even a one semester-hour course in evaluation. Lazar-Morrison et al. (1980) came to the conclusion after reviewing college program requirements that teacher training in measurement methods was minimal, and requirements for formal course work in the area were vague at best or usually nonexistent. From their review, Rudman et al. (1980) concluded that teachers were not well prepared for constructing their own tests or for interpreting the results of standardized tests. In the one teacher survey available at the time, Ward (1980) reported that only 29% of teachers had taken a measurement or testing course, which is even lower than Goslin's 1967 figure. This result must be viewed with caution since the survey response rate was very low, and more recent research suggests that this figure is low (see below).

Studies in the U.S. in the 1980's indicated that about 70 to 80% of teachers had some coursework in classroom measurement (Green & Stager, 1986; Gullickson, 1982, 1984a; Newman & Stallings, 1982; Yeh, Herman, & Rudner, 1981). This was related somewhat to geographical region, but was common across teachers for all grade levels. In spite of this, evidence presented by Gullickson (1984a, 1986b), Gullickson and Ellwein (1985), and Newman and Stallings (1982) suggests that teachers do not have many of the concepts fundamental to measurement, particularly those related to technical and quantitative aspects (e.g., percentile ranks, grade equivalents, standard error of measurement), and that they are not adequately prepared for their classroom assessment tasks. Barnes (1985) concluded for both cooperating and student teachers that "teachers approached the evaluation with puzzlement and concern. The driving force was multifaceted: (a) Fear of hurting the feelings of the pupils, (b) an apparent lack of knowledge of evaluation techniques and concepts, and (c) the unresolved conflict of the criteria to be used in evaluating [primarily performance versus effort]" (p. 48). Most importantly, researchers such as Gullickson (1984b, 1986a, b), Gullickson and Ellwein ( 1985), Gullickson and Hopkins ( 1987), and Stiggins and Bridgeford (1985) argued that the measurement coursework offered may not correspond to the needs of teachers. They noted in particular the impracticality to the classroom context of various technical procedures to analyze test results. Measurement coursework does appear to have some effect on teachers' assessment knowledge and practices: for example, Green and Stager (1986) found differences in test use between teachers who had two or more measurement courses and those who either had no measurement course work or only one course.

For many teacher preparation programs in the U.S. measurement coursework is not required. Gullickson (1986a) found that 71% of the 33 colleges offered a separate course on educational measurement, and for three quarters of these the course was required for pre-service teacher education. In colleges where the course was optional typically 25% or less of students took it, and where there was no course apparently educational measurement information is taught in another required course that is required. Then an estimated 58% of teachers would graduate with one measurement course. Lissitz, Schafer, and Wright (1986) (also Schafer & Lissitz, 1987) found that for institutions of the American Association of College Teachers of Education from 49 to 55% of students did not take a measurement course. The percentages of colleges that did require a measurement course for elementary and secondary teaching certification were 51% and 45% respectively for the two levels. They concluded that coverage of basic measurement topics in other courses was not comprehensive and certainly not systematic across various program areas. O'Sullivan and Chalnick (1991) reported that "fewer than a third of the 51 teacher certification agencies required specific course work or enumerate competencies in educational tests and measurement for initial teacher certification" (p. 18). They suggested that because of the lack of preservice course work in assessment it may be more effective to provide the training at the recertification level. However, recently Jett and Schafer (1992) reported from their survey in Maryland that approximately two-thirds of the teachers had earned a college credit in a classroom measurement course.

Boothroyd, McMorris, and Pruzek (1992) administered to Grade 7 and 8 science and mathematics teachers a test of measurement knowledge related to the *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers et al., 1990), and also asked the teachers to identify flawed test items. About half of the teachers had completed a measurement course, although they reported that it had emphasized standardized tests. The teachers obtained an average of 53% of the items correct, and tended to perform better on items related to writing items and test construction and preparation. As with the Newman and Stallings study, they did poorest on technical aspects such as item analysis (29%). They also performed poorly on grading and reporting. The teachers could distinguish flawed from non-flawed items but had difficulty recognizing flaws in items. The researchers concluded "teachers' knowledge of measurement is not sufficient" (Boothroyd, McMorris, & Pruzek, 1992, p. 7). Marso and Pigge (1992) reported an interesting anomaly regarding teachers' measurement skills: apparently there was a negative relationship between supervisors' ratings of teacher skill and competence and the quality of the test items produced.

Teachers in Canada may have less of a measurement background than do their U.S. counterparts. McLean (1985) states of Canadian teachers that "teachers receive little formal training in evaluation (sometimes none at all)" (p.33). Webster (1987) surveyed teachers in a large urban division and found that approximately 42% of the had taken one or more courses in tests and measurement, with only an additional 13% reporting covering these topics as part of other courses. She concluded "systematic preparation in the area of tests and measurement would seem to be quite limited for approximately half of the respondents" (Webster, 1987, p. 8). On the Newman and Stallings (1982) test, which was administered as part of the survey, the mean score was 53%--similar to that obtained Newman and Stallings. Webster (1987) also found that performance was poorest on items relating to more technical aspects of measurement (e.g., standard error), but argued that the knowledge tested in these items was typically covered in an introductory course in tests and measurement.

This researcher administered the Newman and Stallings (1982) test to two groups of practicing teachers who were taking post graduate educational psychology courses at the University of Manitoba. The results of teachers with no measurement courses in their background was 63% of items correct on average (*n*=12), which is somewhat higher than that for the Newman and Stallings sample. The second group of teachers had just completed a course in test theory and construction, and their mean was 86% (*n*=13). The more technical content covered by the test had certainly been taught in the course, and the results indicated that students had learned it. The teachers in the two groups were fairly similar in background and teaching experience. The results suggest that a measurement course can bring about high scores on the test, and that those teachers who had taken a measurement course in the Newman and Stallings and the Webster samples either (a) did not receive instruction on the relevant topics (unlikely), (b) did not learn the material very well, (c) did not retain the learning, or (d) did not care enough to respond very seriously.

Rogers (1990b, also 1991) reviewed measurement courses offered by faculties of education in Canadian universities and described it "as a patchwork of assessment training at the undergraduate level for both elementary and secondary teacher education students" (p. 31). He reported that 25 of the 33 faculties offer at least one undergraduate measurement course, and that 10 faculties require a measurement course of all prospective teachers with an additional five requiring it of elementary students only. The course is an elective for both elementary and secondary students in 10 faculties and for secondary students only in seven faculties. Rogers (1990b) concluded that of the population of prospective teachers in Canada as many as 60% in elementary and 75% in secondary programs **will not** have a course in measurement and evaluation. They may have received some assessment instruction in other education courses, but this instruction would be incomplete and haphazard at best.

Educators in both the U.S. and Canada are clearly concerned that many teachers have not received a preservice course in educational measurement. On the other hand, much of the coursework in educational measurement that teachers have taken does not appear to fit the bill, and teachers with measurement training seem not to have been adequately prepared. It may be that the classroom situation is such that much of what is taught in a typical preservice measurement course is irrelevant (e.g., Gullickson & Ellwein, 1985; Shulman, 1980; Stiggins & Bridgeford, 1985), impractical (e.g., Gullickson, 1985; Stiggins, Conklin, & Bridgeford, 1986), or simply not learned well enough to be useful or to translate into practice (e.g., Carter, 1984, 1986; Green & Stager, 1986; Webster, 1987). It may be the case that teachers with measurement training do not retain the technical material covered in a measurement course, particularly after several years of teaching. This would argue for both preservice training in measurement and continuing inservice professional development of teachers. It may be as O'Sullivan and Chalnick (1991) note, that some recertification training in measurement is necessary for teachers to retain their licenses.

Many educators have responded to the cry for upgrading teachers' measurement skills. There are more introductory measurement textbooks available now than ever, and in some of these a clear attempt has been made to incorporate a wider range of assessment techniques including such things as observation scales, checklists, and affective assessment techniques (e.g., Airasian, 1992; Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Popham, 1990; R. M. Thorndike, Cunningham, R. L. Thorndike, & Hagen, 1991). Other educators are attempting to address teachers in the field by developing inservice programs and materials and self-help packages. For example, Wanous and Mehrens (1981) argued that there was a distinct need for assessment training that integrates assessment with instruction, and produced a teacher training package for use by individuals or groups of practicing teachers. However, Harold Gulliksen of Educational Testing Service (ETS) commented "over the last 35 to 40 years, the quality of instruction of teachers to write good exam items for their classes has probably declined rather than improved" (1986, p. 112). ETS supported a committee to produce a teacher handbook for item writing (Carlson, 1985), and produced a teacher auto-instructional package to improve test construction skills. Stiggins (see 1987b) was in the process of producing instructional materials for preparing teachers to utilize applied performance testing, and he continues to give workshops on the topic. Item banks are being produced by many state and district agencies, some of which are for teacher use. Other sources of testing materials, such as Popham's IOX materials (Popham, 1984) have also been developed for teacher use. The National Council on Measurement in Education regularly publishes instructional modules on topics in measurement in its journal *Educational Measurement: Issues and Practice*.

There are many ways that teachers' assessment skills could be improved, which include preservice and inservice instruction. One of the tasks is to determine what is appropriate teacher preparation in assessment, and to develop coursework and other means of delivery.

## Content Covered in Measurement Textbooks

Stiggins and Bridgeford (1985) compared teacher-reported assessment practices with percentages of pages of coverage in six textbooks commonly used in introductory educational measurement courses, and found considerable discrepancy (summarized in Table 3 below). If the textbooks are an indication of what was taught in measurement courses, teachers in these courses were given some training in construction of paper-and-pencil tests but little if any in construction and use of alternative assessment procedures, procedures which are clearly emphasized by practicing teachers.

This researcher reviewed the content covered in eight educational measurement textbooks: Ahmann and Glock (1981), Ebel and Frisbie (1986), Gronlund (1985), Mehrens and Lehmann (1984), Nitko (1983), Popham (1981), R. L. Thorndike and Hagen (1977), and Wiersma & Jurs (1985). Seven of these textbooks are well recognized in the field, six of which were in a third or later edition, whereas only one was a relative newcomer. (Since this review was conducted more recent editions have been published by Ebel & Frisbie, 1991; Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Popham, 1990; R. M. Thorndike, Cunningham, R. L. Thorndike, & Hagen, 1991; and Wiersma & Jurs,1990). Of the eight textbooks reviewed, four were the same or a later edition of those reviewed by Stiggins and Bridgeford (1985), which are listed in the footnote to Table 3.

**Table 3.** Relative Emphasis Accorded Major Types of Classroom Assessment by Teachers and by Six Commonly Used Measurement Textbooks[a]

| Major assessment category | Emphasis of teachers[b] | Emphasis of textbooks[c] |
|---|---|---|
| Teacher-made objective tests | 34% | 47% |
| Published tests | 19 | 47 |
| Performance assessment | 47 | 6 |

[a] Abstracted from Table 5 in Stiggins and Bridgeford (1985).

[b] Mean percentage weights attributed by teachers to the assessment categories (averaged across assessment purposes and teachers).

[c] Mean approximate percent of pages on topics in the assessment categories (averaged across six introductory textbooks: Ahmann & Glock, 1981; Brown, 1983; Ebel, 1979; Gronlund, 1981; Mehrens & Lehmann, 1984; and Noll, Scannell, & Craig, 1979).

Many of the same topics are covered in all eight books reviewed, although at different levels of complexity and to varying degrees of detail. For example, all eight devote a chapter or a major section of a chapter to the topic of reliability and standard error of measurement, whereas only some deal explicitly and in detail with reliability of criterion-referenced tests. Table 4 below indicates the average number of pages devoted to various topics. The number of pages is an approximate count to the nearest half-page based on the primary topic covered in that half-page, and includes such things as chapter introductions and summaries, illustrations, tables and lists, diagrams, and appendices, but does not include tables of contents, reference lists, indices, chapter questions, or chapter advance organizers. The textbooks were purposely chosen as ones that are likely candidates in either a teacher undergraduate or postgraduate training program, and their emphasis is certainly not on standardized tests, nor is it on the technical issues of measurement (e.g., scaling, reliability, validity). The bulk of the text coverage is related to test construction, scoring and reporting of results, and test refinement.

The relative emphasis devoted to teacher-made objective tests (Topic 4 in Table 4) is 11%, to standardized/published tests (Topics 13 and 14) is 22%, and to performance assessment (Topics 3 and 5) is 12%. A somewhat different picture emerged from that for textbooks given by Stiggins and Bridgeford (1985). The discrepancies are not due entirely to differences in the textbooks reviewed, although variation in emphasis across textbooks is considerable, and Brown (1983) is oriented more to psychological testing than the other four textbooks chosen by this researcher. It is apparent that many of the modern educational measurement textbooks provide coverage of more than objective item formats, standardized testing, and technical issues of measurement. In fact, seven of the textbooks, all except Ebel & Frisbie (1986), address performance assessment in some detail. And Popham (1981) provides no coverage of standardized tests at all.

## Course Content for Teacher Preparation in Educational Assessment

It appears that approximately every 10 years there is a call to arms for educators to rethink the problem of preparing teachers for classroom assessment. This typically included a study of what knowledge and skills measurement specialists considered vital to teacher preservice measurement courses: for example, in their respective decades, there were Noll (1955), Mayo (1964), Goehring (1973), and Gullickson (1986a) (also Gullickson & Hopkins, 1987; Nitko, 1991a; Rogers, 1990b; Stiggins, 1991a). More recently, four professional education associations in the U.S. (Sanders, 1989) have combined their efforts to produce the *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers et al., 1990). These *Standards for Teacher Competence* are intended to guide educators in the preparation and professional development of teachers in classroom assessment, and to enhance the overall quality of assessment in schools.

**Table 4.** Relative Emphasis Given to Various Topics by Eight Measurement Textbooks

| Topic | Mean across eight textbooks: Number of pages | Mean across eight textbooks: Percent of total # of pages |
|---|---|---|
| 1. Introduction, definitions, purposes, issues, trends | 41.5 | 9.6% |
| 2. Preparation for testing: objectives, taxonomies, specifications | 38.6 | 8.9 |
| 3. Essay item construction and scoring | 15.0 | 3.5 |
| 4. Objective items: short answer, multiple choice, matching, T-F | 46.9 | 10.9 |
| 5. Performance assessment, peer appraisal, affective assessment | 35.8 | 8.3 |
| 6. Assembling, administering, scoring tests | 10.4 | 2.4 |
| 7. Item analysis | 12.2 | 2.8 |
| 8. Descriptive statistics and norm referencing | 46.4 | 10.7 |
| 9. Reliability | 23.1 | 5.3 |
| 10. Validity | 21.9 | 5.1 |
| 11. Marking, grading, and reporting results | 23.3 | 5.4 |
| 12. Test taking characteristics and skills | 7.1 | 1.6 |
| 13. Standardized tests: achieve., apt., personality, interest,... | 85.6 | 19.8 |
| 14. Program evaluation, planning school testing programs | 8.6 | 2.0 |
| 15. Topics unique to one or two textbooks (e.g., diagnosis) | 15.5 | 3.6 |
| Mean total | 431.8 | 100% |

Rogers (1990b) reported that the *Standards for Teacher Competence* were considered appropriate to the Canadian context by a number of measurement specialists, practicing teachers, and administrators. There is some justification, then, for using the *Standards for Teacher Competence* to assist in the development of coursework in classroom assessment. They are:

1. Teachers should be skilled in *choosing* assessment methods appropriate for instructional decisions.

2. Teachers should be skilled in *developing* assessment methods appropriate for instructional decisions.

3. Teachers should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.

4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.

6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment.

There are other documents which address specific aspects of the assessment task, and these can also give guidance. For example, the *Code of Fair Testing Practices* (American Educational Research Association et al., 1988) refers to the ethics of test development and use in education and is based on the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1985). Frisbie and Friedman (1987) have identified some 90 standards from this document which are applicable to teachers in the classroom. Of these 42 were deemed directly relevant to teacher-developed tests, whereas the others were related to teachers' standardized test use.

There are a number of recent articles wherein researchers attempted to identify what teacher preparation in assessment might look like, and what the main ingredients would be. Schafer (1991) specified eight content areas that ought to be included in a classroom assessment course, and rationalized them in part from the *Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practices* . These areas could become an outline for the basic units of instruction for the course, and the skills he identified could serve as general objectives for student learning outcomes. The content areas are: (1) Basic concepts and terminology of assessment, (2) uses of assessment, (3) assessment planning and development, (4) interpretation of assessments, (5) description of assessment results, (6) evaluation and improvement of assessments, (7) feedback and grading, and (8) ethics of assessment (Schafer, 1991, pp. 3-6). These content areas are broadly conceived and encompass many of the topics which are typically included in introductory measurement textbooks.

Recent research on classroom assessment practices clearly portrays the importance of informal assessment procedures (e.g., Airasian, 1991; Gullickson, 1985; Salmon-Cox, 1981; Stiggins, 1985, 1986a, b; R. J. Wilson, 1990). Further, the need to develop and use more authentic assessments, assessments which focus on the most important outcomes of schooling for our students and that are structured in ways consonant with these outcomes, has been clearly articulated and justified (e.g., Brandt, 1989; Burstall, 1986; Crooks, 1988; Fairbairn, 1988; Haney & Madaus, 1989; McLean, 1990; Nickerson, 1989b; Shepard, 1989; Wiggins, 1989a, b; Wolf, 1989; Wolf, Bixby, Glenn, & Gardner, 1991)). Preparation of teachers for classroom assessment must acknowledge this. Based on what Gullickson (1984b, 1986b) had earlier reported as most important to teachers, Linn (1990) suggested these topics as a starting point in setting out the content of a classroom assessment course:

1. Planning and constructing classroom tests. This includes the development of a variety of assessment materials and formats that clearly reflect the important learnings in class. But he noted "to be more effective, however, these materials need to integrated with both subject matter content and instruction in methods" (p. 429).

2. The use of nontest evaluation procedures. Assessment procedures other than paper-and-pencil tests are necessary in the classroom. These are often subjective procedures, and include observations of behaviours, oral questioning, and rating of products.

3. The use of assessment results for instructional planning and formative evaluation. Both tests and other forms of assessment must be used to provide formative data. The data must provide information on complex as well as simpler forms of learning, and must be sufficiently detailed to aid in making diagnostic and group instruction decisions.

4. The use assessment results for summative evaluation. Although formative uses of assessment are primary in the classroom it is necessary for teachers to provide students and parents with summative evaluations. Grades and other forms of summative reporting must be understood.

5. Administration and scoring of tests. Evaluation of higher-level learnings using authentic assessment procedures is complex and difficult. Teachers must be taught how to conduct authentic evaluations and to score such things as essays, projects, and other applied performance assessments.

6. General assessment information regarding the selection and use of tests. Although coverage should be elective and limited, teachers should have a general understanding of standardized tests and how they can be used appropriately. This includes ethical principles of test use in the classroom.

7. Principles of measurement. This includes the concepts of reliability and validity and their guiding role in evaluating the quality of evaluative information from all sources and in all contexts.

Based on his analysis of the literature and a review of outlines for assessment courses across Canada, Rogers (1991) suggested these features should be given prominence in developing a teacher preparation course in classroom assessment:

1. Grading and reporting of student progress must be adequately covered.

2. Nontest procedures for assessment must be stressed, particularly for students in primary and elementary education programs. Teachers must be prepared in the full range of assessment procedures.

3. Authentic assessment approaches, including applied performance assessment, must be a part of all teachers' preparation. Teachers must be able to obtain evaluative evidence from actual work samples and from behaviours of students.

4. Computer use related to assessment must be presented. Teachers should have the skills to use computer technology for a variety of assessment-related functions: preparation of materials, maintaining records, quality control.

5. Assessment design that integrates the assessment with the instruction must be developed. This includes identifying the purposes and procedures of the assessment is done and how the data is to be interpreted.

Finally, in a recent document entitled *Principles for Fair Student Assessment Practices for Education in Canada* (1993), five principles including some 34 guidelines are stated which relate directly to teachers' classroom assessment practices. These incorporate many of the standards and assessment topics discussed above, but provide further specification of what is important for prospective teachers to know and be able to do. The 34 principles are too extensive to present here, but the five themes in which they are grouped give some indication of their content: (1) developing and choosing methods for assessment, (2) collecting assessment information, (3) judging and scoring student performance, (4) summarizing and interpreting results, and (5) reporting assessment findings.

## Summary

A number of educators over the years have expressed concern regarding the quality of evaluation activities in our schools and, more particularly, of teachers' classroom assessment practices. The Canadian Education Association, motivated by this concern, commissioned a study of student evaluation practices in Canada by McLean (1985). He determined that classroom assessment is more of a craft than a science, and teachers learn this craft through trial and error and experience. He concluded that teacher skills in student evaluation are weak. Stiggins and Bridgeford (1985) reported that three quarters of the U.S. teachers they surveyed indicated concern with their classroom assessment practices. Stiggins, Conklin, and Bridgeford (1986) reviewed the literature on classroom assessment practices and concluded that teacher training in measurement was not adequate and the courses available in teacher training programs did not typically provide teachers with training appropriate to the demands of the classroom. These concerns have led to two main areas of research: classroom assessment practices and teacher preparation for assessment. The two areas of research have been the focus of considerable activity both in the U.S. (e.g., Nitko, 1991a) and in Canada (e.g., J. O. Anderson, 1990; Rogers, 1990b). The *Standards for Teacher Competence in Educational Assessment of Students* (American Federation of Teachers et al., 1990) have been developed. These have been judged to be appropriate to Canadian teachers by a number of educators (Rogers, 1991). More recently, the *Principles for Fair Student Assessment Practices for Education in Canada* (1993) have been published.

By far the major part of assessment in the classroom, for student evaluation as well as other purposes, is conducted by teachers using procedures and materials they develop themselves. Standardized tests and other forms of external assessment materials are no substitute for teacher-made materials. Teachers make little use of standardized tests, and they typically view external assessment materials as less instructionally valid than their own testing materials. The amount of assessment activity that occurs in classroom is estimated to consume from one fifth to one third of in-class instruction time. This includes both formal and informal assessments. Further, it is clear that the quality of classroom assessments generally is poor. One critical finding is that higher cognitive level thinking is typically not assessed in any systematic way. The research findings on classroom assessment practices are discussed in detail above. They are also summarized briefly in recent papers by Marso and Pigge (1992) and by Rogers (1991).

## Purpose and Focus of the Study

The improvement of classroom assessment practices is of paramount importance. To improve the quality of classroom assessment, it is necessary to identify clearly what is happening in classrooms and the factors that are related to and impinge on various practices, such as grade levels and subject areas. We now have substantial information on assessment practices generally, but we must obtain a better idea of what is reasonably possible in the classroom. We also must identify ways to provide better preservice preparation and professional development of teachers in measurement. This preparation should reflect both the ideal in assessment practices and the real situation of teachers in classrooms. To identify instruction for teacher preparation in measurement that meets these needs it is necessary first to identify what teachers' classroom assessment practices may be possible ideally, and then to determine what practices are typical (to obtain a "realistic ideal"). The recommendations derived from this must also be evaluated for their utility and practicality by measurement and curriculum specialists and by practicing teachers.

The goal of this study was to develop an effective instructional component for the preparation and professional development of teachers in the area of measurement and evaluation. The component must be consistent with present theory and recommended practices in classroom assessment, and should encompass the goals and objectives of the school program. The component must also reflect the practical realities faced by teachers in their classrooms since so much of the development in measurement theory has not translated into good teacher testing practices.

## Purposes of the Study

The four major purposes of the study, and how they were approached in the study, are:

1. To describe in detail classroom assessment practices of four experienced teachers in two major subject matter areas, science and social studies. To form a broader basis for the recommendations for preparation of teachers in assessment, this description included a review of the literature on classroom assessment practices.

2. To appraise the quality of classroom assessment practices, as observed in classrooms, based on the descriptions obtained for the first purpose. General criteria for the review of observed classroom assessment practices were obtained from the measurement literature, including educational measurement textbooks and the *Standards for Teacher Competence* (American Federation of Teachers et al., 1990). The categories are applied to the teachers' purposes and practices of classroom assessments, and to their assessment instruments.

3. To specify a recommendations for classroom assessment practices and for the preparation of teachers based on broad principles of good classroom assessment. These were to be defined in terms of principles of assessment, but also were to accommodate the practical realities of the classroom context. It was intended to identify possible structures and procedures for delivery of this instruction in assessment.

4. To evaluate the recommendations for classroom assessment practices in terms of their appropriateness, practicality to the classroom setting, and fidelity to measurement principles. They were to be reviewed by professional educators, including practicing school teachers and administrators, education curriculum and pedagogy specialists, and classroom measurement specialists. The review was to result in a revised structure which was to form the basis for defining a program and procedures for teacher preparation in classroom assessment.

## Focus and Features of the Study

The focus of the study was on preparing prospective teachers in classroom assessment procedures. This meant that the emphasis was on teacher preparation, preservice, rather than on professional development and inservice support. It was necessary, however, at times to discuss the implications for inservice training as some aspects of assessment may be more appropriate to practicing teachers. There may also be teacher needs for retraining and further development in assessment. These are referred to in the discussions.

The first two purposes required observing teachers and collecting information on their assessment knowledge, practices, and skills. The procedures for obtaining the assessment information from teachers and the review and appraisal are reported in Chapter III. The intent of the teacher information was to determine what assessment practices might be reasonably possible, so intensive study of experienced, exemplary, teachers was to be undertaken. This is in the form of case studies of four teachers, two in science and two in social studies.

The third purpose involved formulation of recommendations for teacher preparation in assessment. The procedures for instructing teachers for their work in schools could have many different forms, which depend on the approach taken and the assumptions made. It is assumed that teacher preparation and professional development is generally in the form of formal instruction and other forms of structured experiences. The focus was on developing general guidelines for this instruction and experience, but incorporating fundamental principles of good assessment practices. There is a substantial body of knowledge relating to classroom assessment, but this will feature secondarily to the overall structure and focus of teacher preparation in assessment. The basis for the recommendations and the recommendations are presented in Chapter IV.

Recommendations are useful if they identify important features of the problem. They must also hold potential for application. The recommendations were reviewed by experienced educators to ensure that they identified what is significant, and that which is relevant. The results of this review are presented in Chapter V. The implications of the review for the teacher preparation in assessment are presented in Chapter VI.

# III. THE CASE STUDIES

The purpose of the study was to develop an instructional component for teacher preparation and professional development in assessment which was grounded in actual classroom assessment practices. To provide this grounding it was necessary to determine in some detail the assessment practices of teachers. The first part of this phase of the study was to observe and describe classroom assessment practices. The second was to appraise the quality of these practices against criteria for good assessment.

As noted by Stiggins (1987a), the best way to obtain information on classroom practices is to interview teachers, observe their classrooms, and review their assessment materials. The amount and depth of information sought from each teacher meant that only four were selected for this phase of the study. Given the small number, it was necessary to identify teachers who could be expected to model the best of assessment practices in the field, and thereby obtain a detailed understanding of what is possible. The observations of and interviews with these teachers form the cases studies.

## Organization of the Chapter

This chapter begins with a description of how the teachers were selected, followed by an outline of the case study process. The field observations and interviews were organized according to six aspects of the assessment process identified by Stiggins (1987a), although not all of these were used as headings for topics:

1. Purposes which assessment can serve,
2. Methods or techniques used by teachers to assess achievement,
3. Methods or techniques used by teachers to assess affect,
4. Criteria used by teachers to determine assessment methods,
5. Sources of assessment knowledge used by teachers, and
6. Manner in which teachers use their assessment time.

The results of the case studies were divided into three sections: results of the structured part of the interviews; results of the unstructured parts of the interviews combined with the observations of classroom practices; and results of the analysis of the quality of actual assessment instruments, procedures, and reporting mechanisms.

Factors such as grade levels, subject matter areas, demographic mix, geographic regions, and political jurisdictions have been identified as related to classroom assessment practices (e.g., Gullickson, 1985; Marso & Pigge, 1992). There are also considerable differences in practices from teacher to teacher, and from school to school, within a grade level and a particular district or division. Since the classroom setting has an impact on assessment practices, a number of decisions were made to enable the study to concentrate in detail on a limited arena of teachers and settings. The most general constraint placed on the case studies was to conduct them in a suburban setting of Manitoba. This was a matter of convenience to the investigator, but it meant that the Manitoba programs of studies and curricula formed the basis for determining congruity between classroom assessment practices and program goals and objectives.

## Selection of Teachers for the Case Studies

It was necessary to limit the number of case studies to four teachers so that ample time could be devoted to each teacher selected. To obtain an adequate sample of classroom assessment practices each teacher had to be observed over a length of time that ensured the occurrence of a full range and variety of assessment procedures, and so that a clear understanding of what the characteristics and effects of the assessments were. A rationale for the selection of teachers and for the procedures used in the case studies is given below.

## Rationale for the Teachers to be Studied

In Chapter II it was concluded that grade level and subject matter affected classroom assessment practices. For example, Gullickson (1982, also 1985) reported different assessment practices for grades 3, 7, and 10, with large differences for several evaluation techniques between elementary and secondary teachers and for curricular subject areas of science, social studies, and language arts. Although these differences were not consistent across all assessment procedures or for all studies, the general conclusion was that differences could be expected between teachers at the lower grades and those in the secondary schools. Also differences were found at the secondary level between teachers of mathematics and science and those in the humanities.

Given resource constraints, it was decided to select teachers and classrooms at the grade 7 to 9 levels, located in junior high schools (i.e., schools organized as junior high), and to regular or mainstream program settings. These grade levels have many commonalties in structure and program, and often teachers teach at all three levels, so that this selection allowed for some examination of grade level effects within teacher. Most larger junior high schools are departmentalized and teachers usually teach one or two major subject areas (e.g., mathematics and science). Students typically take all courses. Thus, several subject areas could be identified and all students would take the courses at a grade level. This permitted selecting teachers of different subjects and comparing their assessment practices for essentially the same students, which is not readily possible in high schools or most elementary schools. Furthermore, classroom schedules are typically arranged in fixed time blocks either daily or in four, five, or six day cycles. This made scheduling teacher observations and interviews more convenient.

Teachers of two major subject areas were selected, science and social studies. Usually different teachers teach these courses. Both subjects have substantial content components (facts, concepts, principles, etc.) as well as a process objectives (skills, strategies, etc.), yet teacher assessment practices could be expected to differ for the two areas (e.g., Gullickson, 1985). All students in a typical junior high school take both science and social studies. These subjects were chosen over language arts because of their clear content component, which was not as distinctive a part of the language arts program at these levels. They were selected also because there was much less assessment material (published or commercial, curriculum/text imbedded, or other) available to teachers in science or social studies than there was in either mathematics or language arts. Thus, the potential for teacher-generated assessment in these areas was high.

The programs of study in science and in social studies specify a broad range of cognitive abilities which could be analyzed and categorized using schemes such as those adapted from Bloom's taxonomy (e.g., Bloom, Hastings, & Madaus, 1971) or from Biggs and Collis (1982). This permitted the analysis of teacher assessments into various levels or types of cognitive abilities presumed to be required by the items. Finally, science and particularly social studies include both closed and open aspects of learning, in the language of Biggs and Collis (1982), thereby providing latitude for the full range of assessment formats and techniques.

It was difficult to determine the number of teachers that should be included. Since the study was to be intensive in nature the number could not be large. It was decided that four teachers (two each from science and social studies) could be studied given the time requirements of the investigation. To control for the variety of school contexts, it was advisable to select one science and one social studies teacher from each of two schools, with the added intent of choosing teachers who taught some of the same students.

The teachers were experienced in the subject area. They were also considered exemplary teachers by the school principals and by district administrative supervisors. We know from the literature that teachers' assessment practices are typically not of high quality (e.g., Fleming & Chambers, 1983; Haertel, 1986). It was useful to study experienced, exemplary teachers regarding their assessment practices to obtain an indication of what is possible in classroom assessment, yet not unrealistic to incorporate in the goals for teacher preparation in assessment. This could not be obtained by observing and interviewing randomly selected teachers (typical teachers). It was reasonable to infer that the assessment practices of teachers in general would be of poorer quality than those of exemplary teachers. It was also likely that exemplary teachers would demonstrate a wide range of assessment possibilities, which would provide a better basis for developing and refining the recommendations.

### Characteristics of the Teachers Selected

All four teachers selected were males who had taught in the public school system for more than 16 years, ranging from 17 to 22 years. Three of the four had taught at the grade level at which they were being observed for 14 years or more; the fourth teacher for 2 years, but he had taught at other grade levels in junior high for 4 years. The teachers had considerable experience with the subject area, with from 5 to 22 years teaching predominantly in the subject, and from 2 to 15 years in the subject at the particular grade level. The four teachers all had at least four years of university and a bachelor's degree, and all had taken courses beyond this level.

The teachers indicated familiarity with assessment methods and principles, and two of the four teachers reported that they had taken a course in tests and measurement in their undergraduate university education, but none had recent courses in the area. As one teacher noted regarding his undergraduate teacher training, "I took a course in tests and measurement, and administering standardized tests and things like that, but long, long, ago". They all indicated an interest in curriculum. One science and one social studies teacher had considerable experience in developing curriculum materials, and this science teacher had been involved in developing assessment materials at a provincial level.

## Case Study Procedures

The assistant superintendent of a large urban/suburban school division in Manitoba, with responsibility for program implementation and professional development, was asked to nominate four teachers in the division to participate in the study. The study was discussed thoroughly with him, and the request was then formalized by letter (see Appendix A for a copy of the letter). He identified the teachers in consultation with the principals of several junior high schools. Teachers were selected on the basis of endorsement from the assistant superintendent and the principal, given that they were clear on the nature and purpose of the study, and what was required. As supervisory personnel they could readily identify teachers who were exemplary. The teachers were then approached by the researcher to participate in the study, and all four who were approached agreed. Identities were kept confidential.

A variety of procedures and techniques were possible to obtain information from observations and interviews and to conduct case studies of individuals (e.g., Burgess, 1985; Goetz & LeCompte, 1984; Spradley, 1979). The procedures identified by these and other, writers vary greatly depending on the intent and focus of the case study (e.g., Stenhouse, 1988). The present research addressed a number of questions based on the suggestions of Stiggins (1987a), who outlined the major factors relating to classroom assessments and questions that pertain to each. Thus, the case studies were focused in their intent and structured to provide information related to specific areas, and not ethnographies as these are commonly perceived (e.g., Goetz & LeCompte, 1984).

Following LeCompte and Goetz (1982), procedures were used to help ensure the integrity of the observations. One approach was to describe accurately and thoroughly the phenomena being observed, and to obtain observations in a number of situations or over a sufficient length of time to confirm the significance of the phenomena and to ensure that the interpretations are appropriate. Generalizations were tested through further observation or by comparison with other sources of information. The use of two teachers in each subject area facilitated this.

All observations and interviews were conducted by the researcher. The researcher was introduced to the teachers by the principals, and the project was discussed in detail with each teacher. An observation schedule that was convenient to the teachers was determined, but that assured observation of a variety of activities in the classroom.

Observations spanned several weeks during the months of April and May, with each teacher being observed once or twice per week. This period of time represented the major part of the final reporting period in the school year. Each teacher was observed in several different classes of science or social studies at the grade level where he did most of his teaching. Observations included class periods in which formal and informal assessments occurred, and also included class periods shortly before and directly after formal assessments. Observations were repeated for a particular classroom of students throughout the observation time so that a more complete picture could be obtained of all that might happen in assessment with one group of students. Notes were recorded during the observation period, and some time was set aside immediately after to further describe the assessment-related activities.

Each teacher was interviewed after the observations were complete. The interviewer allowed considerable latitude in the interview, and although direct questions were asked, the observed practices of teachers were also discussed. Structured interviews were conducted with each teacher, with the focus being the six aspects of the assessment process identified by Stiggins (1987a) and outlined above. As part of the interview the teachers were asked to respond to each of these six aspects using a written questionnaire (see Appendix B for the directions to teachers and the response forms). Oral questions were posed based on these six aspects, with further probing according to the categories of response for each aspect, as identified in the questionnaire (see Appendix B). Interviews were recorded on audio tape (with permission of the teachers) so that they could be reviewed and summarized later. The teachers were asked to submit to the researcher samples of all the assessment instruments used in the class being observed for the reporting period of the observations.

## Classroom Observations

Each teacher was observed for five to seven classroom periods during the months of April and May. A classroom period ranged in time from 35 to 50 minutes, and additional time was provided either before or after the period for the teacher to make any comments or talk about the class. Various activities were observed, including direct instruction (lesson presentation, demonstration, etc.), in-class assignments and laboratory activities, preparation and review for testing, actual testing, test return, and review and discussion of tests and assignments. The approaches used varied considerably from teacher to teacher and between subject matter areas. Students were often engaged in reading material and answering questions about the material, the questions either being those developed by the teacher or those accompanying the materials, although the amount of this also varied from one teacher to another.

Much of the work in the classes of both science teachers involved students in laboratory activities. Some were detailed assignments of particular tasks with questions to be answered based on the tasks, an example being the use of a microscope to view a single cell organism and to sketch the structure of this organism. Others were less structured experimental activities, such as testing for the *ph* level of common household cleaners. In the social studies classes much of the classroom work was dependent on teacher presentation and on reading the textbook.

In all the classes tests as well as papers and projects were marked by the teacher, although a few of the in-class and homework assignments were marked in class, either by the students marking their own papers or each others'. Teachers reviewed both the assignments and the tests after they had been marked, usually by selecting certain items and discussing them in class. Students were typically invited to give their answers to these selected items and also to expand upon the answers of other students. Students were usually given feedback on their responses, and encouraged to review their papers and to ask questions regarding the tests and assignments. Student progress was regularly monitored based primarily on the tests and assignments, and this occurred on a daily to weekly basis: for example, often work was assigned and marked on a two-day cycle. Quizzes and tests also occurred frequently, but this varied considerably from topic to topic and from teacher to teacher. Tests were usually returned within a day or two of administration. General problem areas were identified by the teacher (diagnosing group needs) and discussed in class, but much of the onus of overcoming individual difficulties or problems with the material was left to the students.

## Teacher Interviews

After the schedule of observations was completed, each teacher was interviewed on several occasions by the researcher. The interviews were scheduled during teacher preparation periods or after school, and the length of a session varied from one class period to almost two hours. Total interview time for a teacher ranged from two to five hours. The interviews were audio taped, with the approval of the teachers.

In the structured part of the interview, teachers were asked to identify the relative importance they attributed to a number of alternatives for each of the six aspects of the assessment process. This is discussed below in the section on structured responses. Since it was important to learn from the teachers how they viewed assessment and how it fit into their classroom practices, the remainder of the interviews was not formally structured with a fixed set of questions, but questions were posed, and the six aspects were used as a general guide. The results of this part are discussed in the section on interviews with unstructured responses that follows.

Some time was taken by the interviewer initially to describe what was wanted and to ensure that the language was commonly used by both teachers and interviewer. It was impossible to cover all of the factors identified by Stiggins (1987a), and many of them were not directly applicable to the teachers. For example, the four teachers did little or no assessment for the purposes of individual student diagnosis in science and social studies, and so this topic was not pursued to any great depth.

The information from the interviews was organized according to the topics outlined in Stiggins (1987a); the questions he suggested served as a basis for the discussions.

## Teacher Interviews--Structured Responses to Six Aspects of Assessment

Six major aspects of the assessment process were identified by Stiggins (1987a). For each of these, the teachers were asked to rate the relative importance of a number of alternatives. For example, the first aspect was assessment purposes, and ten purposes were listed as alternatives. Teachers were asked to rate the importance of these ten alternatives by distributing 100 points among them (the instructions to the teachers are given in Appendix B). This produced a scale which allowed comparisons of the relative importance to the teachers of the purposes, but that did not permit direct comparisons with other scales since the ratings were relative to one another. The mean number of points was computed for the four teachers; these are reported in Tables 10 to 15 for the six aspects respectively. Teacher responses are discussed below under each of these six headings.

### 1. Importance of Various Purposes

On average, the four teachers identified assigning grades as the most important purpose of the ten listed, 31 points out of 100 (see Table 5). This purpose was followed in importance by evaluating instruction. However, the points awarded for both purposes ranged considerably for the four teachers, and one teacher rated the purposes in this order: evaluating instruction, diagnosing individual needs of students, communicating achievement results, and then assigning grades. Assigning grades was clearly an important aspect of assessment to the teachers, but certainly not the only purpose to teachers, and some teachers saw this as less important than other purposes.

**Table 5.** Relative Importance to the Teachers of Ten Purposes

| | Points out of 100 awarded each purpose | |
|---|---|---|
| Purposes (listed in order of mean importance) | Mean | Range |
| Assigning grades | 30.8 | 10 - 50 |
| Evaluating instruction | 20.3 | 10 - 40 |
| Diagnosing individual needs of students | 11.3 | 0 - 20 |
| Diagnosing group needs | 10.3 | 5 - 15 |
| Communicating achievement expectations | 7.0 | 0 - 20 |
| Controlling and motivating students | 6.8 | 1 - 10 |
| Communicating affective or behavioral expectations | 6.0 | 0 - 20 |
| Grouping for instruction within class | 3.8 | 0 - 10 |
| Identifying students for special services | 1.3 | 0 - 5 |
| Providing test taking experience | 2.3 | 0 - 5 |

Diagnosing individual student needs and group needs received similar average ratings, approximately 10 points, which was somewhat lower than the two highest rated purposes. The range across teachers was also not as great for the diagnostic purposes, although one teacher rated each of them very low. The lowest rated purposes were grouping for instruction, identifying students for special services, and providing test taking experience. This supports the understanding that teachers at the junior high level typically do not use assessment to identify groups of students within a class for whom they provide systematically different instruction, although one of the teachers mentioned in the interviews that he had done this in the past. Both schools provided special services for students with identified difficulties, and the teachers noted that they may refer students for this assistance, but this was more often done based on performance in language arts and mathematics. All four teachers noted in the interviews, and with some vehemence, that teaching test taking skills was not a good reason for giving tests: they suggested that there were better vehicles for developing these skills. Two of the teachers indicated that preparing students for high school tests is the subject of ongoing debates in the school.

These results corroborated the findings of Gullickson (1982), who reported that at the grade 7 level teachers rated the importance of teacher-made tests to student grades 2.7 on a 0 to 4 scale, but gave similar ratings to instructional feedback (2.6) and evaluation of instruction (2.6). Lower ratings were given to such purposes as motivating students (2.2) and assessment of attitudes (1.4). Webster (1987) concluded that classroom assessment serves two main purposes: provide feedback for instructional planning and provide information about individual students' skills and capabilities. The latter purpose is broad enough to include both diagnosis and grading, so it appears that the results were in keeping with those of Webster (1987). These results also agreed with those of R. J. Wilson (1989) who reported that most classroom assessments, approximately 80%, were used for marks at the elementary and secondary school levels, but frequently also for checking learning progress and diagnosis.

## 2. Importance of Various Methods to Assess Achievement

The teachers identified teacher-developed paper and pencil tests as the most important method to assess achievement, 33 points out of 100 (see Table 6). Although there was considerable range among the teachers, 20 to 60 points, all four rated this method either highest or second highest of the 11 given. This method was followed in importance by regular homework assignments, 15 points, and performance assessments, 14 points. There was considerable variation among teachers for these two methods, but in all cases they were rated no lower than fourth in importance. Text-embedded tests were rated next in importance on average, 10 points, followed by assessment of reasoning skills, 9 points. These two methods were not consistently rated by the teachers, with some teachers rating each 0 or 5 points and one teacher rating each 30 points. There was greater consistency for the next two methods, group assessment and oral questioning, all teachers rating them 15 points or lower. The lowest ratings were obtained, in declining order, for student self ratings, opinions of other teachers, student peer ratings, and standardized tests, which received ratings very close to 0. These methods are rarely used by the teachers and this is reflected in their low ratings.

Previous studies have also noted the considerable variation in assessment practices across teachers of different grade levels and subject areas, as well as among teachers teaching similar courses (Bateson, 1990; Lortie, 1975; Lazar-Morrison et al., 1980; Fennessy, 1982; Gullickson, 1982, 1985; Stiggins & Bridgeford, 1985; Webster, 1987; R. J. Wilson, 1990). Standardized tests of any kind received very little interest from teachers; this was reported by all studies in the literature.

**Table 6.** Relative Importance to the Teachers of Eleven Achievement Assessment Methods

| Achievement assessment methods (listed in order of mean importance) | Points out of 100 awarded each method Mean | Range |
|---|---|---|
| Teacher-developed paper and pencil tests and quizzes | 32.8 | 20 - 60 |
| Regular homework assignments | 15.0 | 5 - 30 |
| Performance assessments | 14.0 | 10 - 21 |
| Text-embedded paper and pencil tests and quizzes | 9.5 | 0 - 30 |
| Assessment of reasoning skills | 8.8 | 5 - 15 |
| Group assessment methods (also projects in pairs) | 7.5 | 5 - 10 |
| Oral questioning strategies (and interviews) | 5.8 | 0 - 15 |
| Opinions of other teachers (other than grading) | 1.8 | 0 - 5 |
| Student **self** ratings | 2.5 | 0 - 10 |
| Standardized tests (school, division, province, etc.) | 1.3 | 0 - 5 |
| Student **peer** ratings | 1.3 | 0 - 5 |

Gullickson (1985) reported that at the grade seven level objective teacher-made tests were rated as having the greatest role in evaluation when compared to the other ten methods he listed: 2.5 on a 0 to 3 scale. This was followed by student papers and notebooks, 2.3, and class discussion, 2.0. The remaining eight methods received ratings below 1.5. These results corresponded with the case study findings for paper and pencil tests and homework assignments, but seemed to differ in the performance assessment category which was rated fairly highly in the present study (although there was no clearly parallel category in the Gullickson study).

Stiggins and Bridgeford (1985) reported for grade eight teachers that the highest level of "comfortable use" was for "spontaneous performance assessment", 84% of teachers, followed by teacher-made objective tests, 68%, and "structured performance assessment", 62%. Amount of use was not clear, but the results implied that observation, which would fall under the spontaneous performance assessment umbrella, was most comfortable to teachers. Teachers in the Webster (1987) survey reported oral interview and work sample as being the most frequently used, followed by observation, and teacher-made tests. Teacher-made tests were reported to receive higher frequency use at the higher grade levels, however. This was supported by Webster's interviews: teachers tended to put the most faith in their everyday observations of students, in assessment of student work assignments, and in the results of their own testing. The other methods were much less frequently used, for example, self assessment and rating scales.

R. J. Wilson (1990) reported that teachers emphasized teacher-made tests, and these were more important than other forms of assessment. Bateson (1990) indicated that grade 7 science teachers rated teacher-made objective tests as "very important" most frequently for deriving final evaluation, followed by projects and lab write-ups.

## 3. Importance of Various Methods to Assess Affect

Three of the four teachers rated the importance of eight methods for the assessment of students' affective traits (one teacher indicated in the interviews that he did little of this kind of assessment and did not think his response would be helpful). The first method was clearly the most important, observing individual students was given the highest average rating, 38 points out of 100 (see Table 7). This was followed by observing group interactions and using questionnaires with 20 points each. Although ratings ranged considerably for these methods, their averages were well above those of the remaining five, which had average ratings from 3 to 7 points. Individual teachers varied somewhat in their emphasis, but the pattern was apparent, and observation was favoured.

**Table 7.** Relative Importance to Three of the Teachers of Eight Affective Assessment Methods

| Achievement assessment methods (listed in order of mean importance) | Points out of 100 awarded each method | |
|---|---|---|
| | Mean | Range |
| Observing individual students | 38.3 | 25 - 50 |
| Observing group interactions | 20.0 | 5 - 30 |
| Using questionnaires | 20 0 | 10 - 30 |
| Opinions of other teachers | 6.7 | 0 - 15 |
| Opinions of parents, guardians | 5.0 | 0 - 10 |
| Using interviews | 3.3 | 0 - 5 |
| Opinions of other students | 3.3 | 0 - 5 |
| Past student records | 3.3 | 0 - 5 |

Few other studies had separately identified assessment practices in the affective arena, but generally observation has been a strongly preferred approach to assessments which can readily include affective components--other than teacher-made tests of achievement and work assignments (e.g., Webster, 1987). Teachers readily acknowledged the importance of the affective domain in education, particularly when interviewed (e.g., Salmon-Cox, 1981; Webster, 1987), but did not as readily identify this area as one which they assess other than by informal observation. This appeared to be true of secondary teachers more than elementary: for example, Gullickson (1985) reported elementary teachers rating the role in assessment of citizenship behavior as 2.1 in importance compared to objective teacher-made tests at 1.9, whereas junior high teachers awarded corresponding ratings of 1.4 and 2.5. On the other hand, Webster (1987) reported that teachers across grade levels did not significantly differ in their ratings of involvement in various areas, but that most teachers rated their highest involvement was in the assessment of achievement: 88% of teachers rated it 4, on a 1 to 4 scale, as opposed to work habits, 73%, social attitudes, 50%, and other affective traits less than 50%. On the same survey 29% of teachers reported that cognitive items were used on their tests once or twice a month and 63% said on most tests, whereas the figures were 42% and 14% of teachers respectively for affective items, and 14% and 12% respectively for psychomotor items.

## 4. Importance of Various Criteria for Selection of Assessment Methods

The teachers identified the most important criteria they use in selecting particular assessment methods as "assessment results fit purpose" (i.e., data obtained fit the particular needs for the assessment, such as grading), 21 points out of 100 on average, and "method matches intended outcomes" (method appears to capture that which was taught), 20 points (see Table 8). The "origin of assessment" (the source where the teacher obtained the assessment) achieved the next highest average rating, 12 points, but this was primarily due to one high rating of 30 points.

**Table 8.** Relative Importance to the Teachers of Nine Criteria for Selection of Assessment Methods

| Selection criteria<br>(listed in order of mean importance) | Points out of 100 awarded each criterion<br>Mean | <br>Range |
|---|---|---|
| Assessment results fit purpose | 21.3 | 10 - 30 |
| Method matches intended outcomes | 20.0 | 10 - 30 |
| Origin of assessment | 11.5 | 1 - 30 |
| Ease of development | 9.8 | 9 - 10 |
| Ease of scoring | 9.8 | 9 - 10 |
| Degree of objectivity | 9.8 | 5 - 15 |
| Applicability to measuring higher order thinking skills | 8.8 | 5 - 15 |
| Time required to administer | 5.0 | 5 - 5 |
| Effective control of cheating | 4.3 | 0 - 10 |

Practical considerations of ease in development and scoring, administration time, objectivity, and control of cheating were not rated highly, from 4 to 10 points (see Table 8). This suggested that these concerns were secondary to that of deciding what assessment was necessary and how could the assessment be designed to fit the intended learnings. The fact that "applicability to higher order skills" was rated 9 points seems to reflect the data from the teacher-made tests that a relatively small proportion of the testing tapped higher cognitive levels of thinking (i.e., beyond the comprehension level of Bloom's taxonomy). Although the teachers differed somewhat in their views, none of them attributed more than 15% to this criterion. In the interviews the teachers stated it was important to assess higher-level thinking, but that the amount of this kind of testing should be limited for students at these grade levels (again, the teachers varied considerably on how much testing should be at higher levels). The actual test documents revealed that only about 10% of marks on tests could be considered to be based on higher level items, items beyond the comprehension level of Bloom's taxonomy.

The earlier findings of Fleming and Chambers (1983) were that very little of the testing material they analyzed required higher-level thinking on the part of the students. Carter's (1984) evidence suggested teachers may have difficulty in recognizing which items assess interpretive reading skills, items at a higher cognitive level, and that these kinds of items are exceedingly time-consuming for teachers to construct. Certainly, evidence exists that higher-level thinking skills were not commonly assessed by teachers (Stiggins, Griswold, & Wikelund, 1989; Webster, 1987). It appeared that unless teachers assess at the higher levels as part of their nontest assessment activities (performance assessment, etc.) it was likely that very little of assessment is beyond Bloom's comprehension level. Teachers put little emphasis on the criterion of higher-level thinking in their search for assessment materials, suggesting that these kinds of items were not actively sought nor developed.

## 5. The Importance of Various Sources of Assessment Knowledge

Perhaps the clearest data were obtained for this aspect of the assessment process. The teachers consistently reported "own classroom experience" as the main source: 49 points out of 100 (range, 40 to 60). The sources "ideas and suggestions from colleagues", "guidebooks accompanying texts", and "preservice and graduate teacher training" were given 14, 13, and 11 points respectively (see Table 9). And "inservice training programs" and "professional literature" played almost no role: 8 and 6 points respectively (see Table 9). It appeared that teachers typically did not obtain information from training programs or from professional printed material, and these are the vehicles which are typically used for teacher upgrading.

Given that there are problems with teachers' assessment practices this information is suggestive of ways to approach professional development of teachers. If these data are representative of teachers today, either appropriate inservice and other forms of professional development in assessment are not available to teachers or it is futile to provide inservice workshops on assessment in the typical format. Certainly the one-shot workshop, without some support and follow-up, was recognized as ineffective generally for bringing about teacher change (Cousins & Leithwood, 1986; Guskey, 1986). Furthermore, providing teachers with assessment information via print materials in professional journals does not appear fruitful. Other approaches need to be designed to facilitate teacher development in assessment. With effort being given to defining assessment skills of teachers, such as in the *Standards for teacher competence* (American Federation of Teachers et al., 1990), and to defining the knowledge necessary for a teacher preparation course in classroom assessment, it should be possible to design appropriate professional development activity for practicing teachers (e.g., Stiggins, 1991).

**Table 9.** Relative Importance to the Teachers of Six Sources of Assessment Knowledge

| Sources of assessment knowledge (listed in order of mean importance) | Points out of 100 awarded each source | |
|---|---|---|
| | Mean | Range |
| Own classroom experience | 48.8 | 40 - 60 |
| Ideas and suggestions from colleagues | 13.8 | 5 - 30 |
| Guidebooks accompanying texts | 13.3 | 5 - 20 |
| Preservice and graduate teacher training | 11.3 | 0 - 30 |
| Inservice training programs | 7.5 | 5 - 10 |
| Readings from professional literature | 5.5 | 2 - 10 |

## 6. Manner in which Teachers Allocate Their Assessment Time

It has been well established in the literature that assessment consumes a significant amount of classroom time as well as outside of class teacher time. The teachers indicated three activities as requiring near equal amounts of time: "developing own assessments", "scoring assessments", and "providing feedback" were allocated 21, 21, and 19 percentage points (of time) out of a total of 100 (see Table 10). "Administering assessments" accounted for 15 points, and "reviewing and selecting assessments" for 13. Although there was considerable individual variation, this suggests that the teachers probably spent more time in preparing assessments, in scoring student responses, and in providing feedback than in administering them to students. Assessment preparation and marking were mostly done outside of classroom time, so it appears that teachers spend more time out of class on assessment related activities than on administering assessments and going over the results with students. If the estimate from the literature is accurate, that at least gioven more than 20% of in-class time is taken up with assessment activities including recording results and posttest analysis, then teachers may spend the equivalent of six hours per week on preparing assessments and marking them.

Recording results was given 5 points indicating that this did not consume much of a teacher's time relative to the other assessment activities. Finally, evaluating assessment quality was given 6 points, which suggests that relatively less effort is given to reviewing the assessment in retrospect. This would lend support to the concern evinced by Gullickson (1982) and Gullickson & Ellwein (1985) that teachers neither knew nor used test analysis procedures to improve their tests. The teachers commented in the interviews that the extra time required for test analysis produced yet another burden on their schedule, and this was at a time when interest was primarily in preparing the next section of material. Given this, careful review of test results and possibly some further analysis, which would be necessary for individual student diagnosis, would also be unlikely.

**Table 10.** Relative Amount of Time Teachers Give to Seven Tasks in Assessment

| Uses of assessment time (listed in order of mean time) | Points out of 100 awarded each use: Mean | Range |
|---|---|---|
| Developing own assessments | 21.3 | 10 - 35 |
| Scoring assessments | 21.3 | 10 - 40 |
| Providing feedback | 18.8 | 15 - 20 |
| Administering assessments | 15.0 | 10 - 20 |
| Reviewing and selecting assessments | 12.5 | 10 - 20 |
| Evaluating assessment quality | 6.3 | 5 - 10 |
| Recording results | 5.0 | 5 - 5 |

## Teacher Interviews--Unstructured Responses and Observations

The six factors and accompanying questions specified by Stiggins (1987a) provided a scheme for the presentation of the information that was obtained from the nonstructured portion of the interviews (see section at the beginning of this chapter). Some of the information was elicited from the teachers by directq questions, but much of it involved description and subjective interpretation by the researcher of the teachers' comments and responses during the interview. This was combined with interpretations based on the classroom observations. The first four of the six factors are discussed in turn below. The last two did not warrant discussion, but an additional topic was included: teachers' assessment of ability.

### 1. Assessment Purposes

There were ten purposes for classroom assessment outlined earlier (see Table 5). Each is discussed in detail.

**Diagnosing individual needs of students**. The teachers appeared to understand individual differences of students, and that certain students would do better than others at some tasks. One of the teachers reported that he asks each student to help identify her/his individual strengths and weaknesses, and uses this in part to help the students focus their effort. One teacher notes that social studies is different from such subjects as mathematics and science and thus presents an opportunity for students who find math and science difficult to do well in social studies. However, there was little evidence that the teachers did individual diagnosis in either science or social studies. None of the four designed their assessment procedures such that the results could be used for effective diagnosis (i.e., specification of objectives, subtest structure, graduated difficulty, repeated assessment of topics which posed difficulties, etc.). They also did not mention these as possibilities. They mentioned that diagnosis is important and useful, but that it was done much more commonly in mathematics and language arts (in fact, there is resource support in both schools for students who are perceived to be having learning difficulties). The teachers did note informally questions that pose difficulty for students and tended to follow these with directed discussion and practice in class; however, this is more group than individual diagnosis.

The teachers noted that the material in science and social studies that they teach is not typically structured to facilitate diagnosis, and appropriate testing materials were not readily available. But the case could be made that process skills, such as hypothesizing and inferring or safety in the lab, could be tested and problems diagnosed for individual students. The load of 20 students in some of the classes would permit this, but other classes had as many as 35 students, making it more difficult. All four teachers taught the same course to several classes, which should amortize the effort of developing assessment materials. The teachers all used questions that required student written responses, and argued that this helped them to determine what the student knew and what kinds of errors were being made. This tends to increase the marking load, and all four teachers noted that marking was very time consuming for them.

Diagnosing group needs of students. The teachers indicated that group diagnosis was done in a general way in their science and social studies classes. They looked for topics and items where students had difficulties, and these would be selected for review and possibly for reteaching. As with individual diagnosis, tests were not designed specifically to facilitate diagnosis although mention was made of the use of various item types to obtain different kinds of information. One teacher did use pretesting to help determine student understanding, and on the basis of this adjusted instruction.

The teachers did not indicate how assessment results would be presented for diagnosis, and it appeared that any diagnosis was based on informal observations of students' work. Since the teachers marked most of the assignments and tests they had ample opportunity to observe what the students were actually doing. All four teachers thought that group diagnosis was only of limited value and that it was less important in these subject areas than in mathematics and language arts. They did not group students in class and provide differentiated instruction in a systematic way, but they did on several occasions informally point to problem areas and help individual students with them.

Assigning grades. It was reported earlier that assigning grades was rated as the most important purpose of assessment, which may be related to the fact that grades and reporting to parents were a necessary part of teaching in the two schools. All four teachers indicated that they used a variety of assessment methods to obtain student grades, including quizzes and tests, in-class and homework assignments, papers, projects, labs, and behavior ratings. The example of a grading scheme for a unit of instruction given in Figure 1 below is based on information from one of the teachers; it gives some indication of what was included for student grading purposes. However, the weights attached to the assessment methods varied across teachers and also across topics.

For these teachers there were usually three components to assessment for grading: quizzes and tests, assignments, and special papers or projects. Other data sources may well have been included in some reporting period, depending on the topic and content being taught, but these three components formed the basis for most of students' marks. The teachers argued that it was important to base grades on both every day work and on tests, thereby including some weighting for endeavor or effort on the part of the students, as this appears in the quality with which they do the assignments. The teachers often provided behavioural and affective assessment as a part of the reporting but usually independently of the grade. This was typically a general assessment of such things as effort, cooperativeness, and willingness, but could also be very specific in that certain behaviors may be noted: "assignment x not completed", "disrupts other in the class", or "works hard and is doing well".

**Figure 1.** Typical Scheme Teachers Used for Grading Students on a Unit of Instruction

| | Marks allotted | | Percentage of total unit marks | |
|---|---|---|---|---|
| Paper and pencil tests | | | | |
| Quiz 1 (one chapter of text) | 10 marks | | 6% | |
| Quiz 2 (two chapters of text) | 15 | | 9 | |
| Unit test | 30 | | 19 | |
| Total marks for tests | | 55 | | 34% |
| Notebook, including in-class assignments | | 45 | | 28 |
| Lab reports | | 20 | | 13 |
| Project | | 40 | | 25 |
| Total Marks | | 160 | | 100% |

Grades and reporting were part of the school division's policy but the procedures and format were the school's responsibility. The actual content that would form the basis of student grades in a course was the jurisdiction of the teacher, but in both schools grade level teams of teachers, and subject area teachers, collaborated to establish what generally they would teach and how they would grade students. Both schools were reviewing their reporting system. One school had established an elaborate procedure that required considerable amounts of teacher time, and teachers had not yet decided whether this was what they wanted. Teachers in the other school were dissatisfied with the lack of detail on their report form and were redrafting the format. All four teachers viewed grading as an important aspect of assessment and were concerned that it be done as comprehensively as possible. Both schools had teachers involved in reviewing the grading and reporting procedures; one school had an active committee of teachers and administrators.

**Grouping for instruction**. None of the four teachers grouped students for differentiated instruction based on the results of assessments. This corresponded with the low rating they gave to this purpose. Grouping of students was frequently done for purposes of lab work or conducting projects, and all four teachers typically allowed students to choose their own partners and groups. Ability grouping within a class simply was not done, and in both schools there was no among-class ability grouping, no streaming. One teacher noted that he had previously grouped his mathematics classes on the basis of ability in conjunction with another teacher, but argued that the present timetable and classroom structure made this difficult. Another teacher suggested that class sizes (of 30 to 35) worked against this kind of differentiated instruction within a class.

Both science teachers agreed that grouping students would make sense in science, particularly with respect to some of the more complex skills in science. One teacher used individual projects in science as a means of permitting students to develop particular topics, a method to provide enrichment but also to promote learning from different settings, including outside of the classroom (e.g., public and college libraries, medical clinics, and outdoors). He encouraged students to opt for projects which took them afield and approximately one third of his students chose this rather than a more conventional paper. However, decisions regarding what was done in these projects and which students were involved was based on factors such as the maturity of the student and ability to work independently, although assessment results entered the picture of what students' capabilities were.

The teachers indicated using tests early in the school term to obtain an impression of students' level of knowledge, but as Salmon-Cox (1981) noted, as the year progresses this impression becomes dependent on a host of factors and on information which is primarily obtained through observations and results of classroom assignments. Yeh, Herman, and Rudner (1981) came to a similar conclusion.

**Identifying students for special services**. All four teachers identified students for special services; these services consisted of a special remedial classroom in one school and classroom aides in the other. No procedures existed structurally in these schools for providing special services for advanced students. Although most of the remedial support was invoked on the basis of student difficulties in mathematics and language arts, teachers did identify students for assistance in science and in social studies classes. Teachers in one school indicated that a student's performance in all classes and courses was used as the basis for determining whether the student should receive special services and be given a special program in a remedial classroom for part of day. It appears that although the teachers rated this purpose very low, they did identify students for special services.

The teachers indicated that they used assessment as part of the evidence for identifying students and that it was important in this regard, but only insofar as these students scored poorly on the assessments they set for their courses. That is, no specific assessments were developed or selected for this purpose, although one teacher did indicate that he thought standardized achievement tests could be useful in this regard.

**Controlling and motivating students**. All four teachers indicated that they used assessments, and tests in particular, to motivate student learning, but more to direct and focus students' efforts than to use them as threats. The teachers perceived assessment as an important facet of the learning process: part of this was to maintain student motivation, but also to keep them aware of where they are in their learning and what they have to do. One teacher noted that assessment can aid students in their work and give them a sense of accomplishment; he specifically developed this reinforcing aspect of assessments by preparing the students (e.g., teaching them how to answer questions, reviewing, etc.) and making them fully aware of how they were doing. Another teacher used assessment to emphasize and encourage certain skills, and, as an example, stated to students "I'm going to be marking your accuracy, your safety practices, and how careful you are with chemicals". He informed students that these things "count and are part of doing well in the laboratory". All four teachers stated emphatically that assessment, particularly tests, should not be used as punishment, rather as motivators, and to control students' behavior by focusing their effort.

**Evaluating instruction**. There was little evidence that the teachers regularly evaluated the effectiveness of their instruction in an explicit and thorough manner. Assessment that is ongoing in the classroom provided some information for this purpose, but the purpose was secondary to other purposes. Although the teachers identified this as significant in their ratings of purposes, only two of the teachers invoked specific procedures for this. One teacher indicated that he tried different instructional approaches from year to year and from class to class, and used assessment to help determine effectiveness. Another teacher administered questionnaires to students to request their input into what was happening in class. Information from these questionnaires was used to help determine if certain teacher practices and classroom activities were perceived as effective by the students. Questions that often appeared on the questionnaires were: "how well did I prepare you for the test?", or "how well do you think you learned the material?". This teacher also asked students to provide direct feedback on his tests regarding aspects such as the format and content.

The four teachers recognized that certain classroom activities were more effective than others for the objectives they were trying to accomplish. Each teacher could identify different approaches that he had used in his classes, such as group discussions or oral presentations, and whether these were believed to be effective. But only one teacher indicated making between-class comparisons on assessment results, although these were also used to help inform students of how they were doing and to motivate students.

One teacher stated that as he gained experience teaching, his interest in evaluating instruction, in monitoring his teaching, increased: "the more time spent in the classroom the more important I think it is". But the typical assessments conducted in class appeared to be more for other purposes, such as for providing feedback to students or for grading, than specifically to evaluate instructional strategies.

**Communicating achievement expectations**. The teachers clearly felt that assessment should not be the key source from which students are to identify the goals and objectives of the course. These, they believed, should be expressed clearly at the beginning of the term and reinforced periodically throughout the year. The common view was that assessments form part of the task of reinforcing what is important and what students are to learn: for example, one teacher noted "tests inform students retrospectively what is to be learned, what they know and don't know and what to strive for". This purpose was rated lower than assigning grades, diagnosing individual and group needs, and evaluating instruction.

Teachers agreed that it is important to inform students of expectations, and suggested that a better approach than to use assessments was to provide students with practice assessments, to train them to recognize what is important and to learn the material prior to actual assessment. Most teachers have expectations clearly beyond the specific content being taught, such as skills in thinking, in reading, in writing; skills in processes like careful observation, recording and displaying information; attitudes and practices like safety, care, and cleanliness in the laboratory, respect for others. The four teachers believed that these too should be expressed openly and in advance, but indicated that assessment of them and setting of standards certainly could reinforce their importance. The fact that some skill is being assessed usually leads to a clearer indication of what the standard of acceptable, or good, performance is--this was recognized by the teachers. One teacher summed up his view of assessment this way: "there should be no surprises".

**Communicating affective or behavioral expectations**. Teachers gave this purpose much the same rating as they did the previous one, but there was some indication that teachers were not as clear about what this purpose meant and spent more time talking about their expectations of students. There was limited assessment in the affective domain generally, other than by means of informal observations and anecdotal recordings. The teachers recognized that affective traits and psychomotor skills are important, and that they could not be assessed in the typical way that achievement is assessed. They indicated some alternative procedures that could make assessment in this domain more systematic, such as scales for the rating of behavior, checklists, and observation schedules. As with the achievement area, however, assessment was not considered as the best vehicle for communicating objectives of an affective and behavioral nature: this, they thought, ought to be done explicitly and in advance through instruction, guidance, and discussion rather than assessment.

One teacher noted that although affective objectives are important, assessment was by way of student achievement; task completion was the important affective criterion and this was clearly stated in advance. Another teacher did use direct assessment of affective objectives, such as lab safety, cooperation, and cleanliness, and argued that this kind of assessment did clarify the criteria and make them tangible for students. On an ongoing basis in all classrooms students were informed if they did not meet certain criteria of behavior: for example, sticking to the task, listening to the teacher and to other students, and conducting themselves in a safe manner.

**Providing test taking experience**. This potential purpose of assessment sparked the strongest response. All four teachers rated it of little importance, and stated emphatically that it was not a good reason for assessment. They felt that there were much better ways of preparing students for tests, direct instruction being one. Teachers in both schools indicated that there was an ongoing debate about this in their school, suggesting that it is an important issue to educators. The teachers agreed that students must learn how to take tests, and that there are students who are unable to perform at their best on tests. They believed some students are not able to understand the test items, to handle the stress, or to write clearly in a test situation; they must be taught to do this and given experience in answering a variety of item types, but before the tests are administered.

## 2. Methods Used by Teachers for Assessing Achievement

There were eleven general categories of assessment methods used by teachers in their assessment of student achievement, and each was discussed with the teachers during the interviews. These discussions are reported below (see also Table 6 above).

**Teacher-developed paper and pencil tests and quizzes**. All four teachers regularly used paper and pencil tests. This category of assessment methods was rated as most important, on average, of the eleven listed for the assessment of achievement. All four teachers indicated considerable familiarity with test development procedures, such as how to construct a multiple-choice item. However, none of them indicated that they used anything like a test blueprint (or table of specifications) or behavioral objectives. All stated that they had a general plan at the beginning of the year indicating approximate length of topics and units, mark allocations, and what format the assessments were likely to take. Further, they specified in some detail at the beginning of each unit how students would be assessed (e.g., the tests and quizzes, assignments, papers, etc.); an example of a mark breakdown for one unit is given in Figure 1 above.

The content of a test was typically not spelled out in great detail at the beginning of teaching a unit, but the teachers described what is to be on the tests and quizzes shortly in advance of administration, and usually provided some preparation and review time. Assignments and papers were given in written form, sometimes on the blackboard, and were discussed with the students at the time they are assigned.

The teachers could identify some of the characteristics of good test questions (such as clear scoring guides for supply-type questions). They recognized some of the problems with particular item types (such as marker subjectivity in long-answer questions), and possible ambiguity in the wording of questions. It was not clear whether the teachers clearly understood the techniques for effective item development, and from the analysis of their tests, it was apparent that there was considerable room for improvement. Two teachers indicated familiarity with techniques of item analysis based on student responses, but none of the teachers actually conducted a formal item analysis. All four teachers informally reviewed the test and specific items after it was administered, and made a particular effort to do so if they planned to use the test again. The teachers asked that students carefully look over tests and assignments that had been marked, to check for errors in mark accumulation and so forth. The teachers were open to questions by students about a mark on a particular item, and to criticisms of the items more generally. However, there was little systematic peer review of tests although the teachers readily shared their tests with others and also worked together with others in the development of assessment materials.

The teachers were aware that much of what is important to learn cannot readily be tested using paper and pencil tests, and this was one of the reasons they gave for including two or three different formats as part of the assessment for a unit (usually assignments and papers). One of the teachers stated that he was moving more towards:

- Test items which require longer written responses, to give the teacher a better opportunity to determine where the student may be having trouble;
- Assignments that include oral presentation components (at least as an option), so that students learn to respond both orally and in writing, and also since some students cannot express themselves as readily in writing; and
- Questions that require a variety of tables, graphs, or diagrams as part of the response, since these techniques are an important means of communicating information particularly in the physical and social sciences.

**Text-embedded paper and pencil tests and quizzes**. The textual materials used by the four teachers typically provided accompanying questions and assignments. Two of the teachers made considerable use of these embedded or associated questions as part of daily work assignments, and these and similar questions often appeared on the teachers' tests. However, most of the tests used by the teachers were constructed by the teachers themselves (confirming a point made by previous researchers). One of the teachers noted that many of the questions in the textual materials required only simple cognitive functioning to answer (e.g., listing facts, or repeating information presented in the text), and therefore of limited value for many of the goals of the school program. This was also a point made by Stiggins (1986b). Further, one teacher noted that assessment items from textbooks may not correspond with what he emphasized in the class, and therefore were not directly applicable. However, the teachers indicated that they adapted and modified materials from a number of sources, and that good use of materials can be made. The teachers comments corresponded with their rating of the importance of this source as being not very high.

**Performance assessment**. Performance assessment was identified to the teachers as direct observation and professional judgment of student behavior and products in settings designed to demonstrate student proficiency of the actual skill, rather than a proxy measure of it (this includes structured and spontaneous performance assessment). This type of assessment was rated as quite important by the teachers (ranked in importance directly after teacher-made tests and assignments). There were some differences between science and social studies teachers as to what types of things were being assessed. The science teachers assessed students' performance in the laboratory; this was primarily in terms of what the students produced (their laboratory experiments and reports) but also included their actual behavior in the lab (safety practices, etc.). The social studies teachers assessed such things as students' ability to conduct library research and to use research information by requesting research essays of their students. Both science and social studies teachers were also concerned with such traits as accurate observation skills, clear reporting, ability to display information (e.g., diagrams and graphs), and use of written language (emphasis often on syntax and grammar). The language skills were commonly assessed as part of the product, the report or paper.

The performance assessments were typically well defined, with clear directions to students. Most also contained mark allocations for aspects of the report. The teachers usually provided students with structured settings and well defined tasks for exhibiting their skills. There was no evidence that the teachers analyzed the quality of their performance assessments, although they were all conscious of the problem of subjectivity in the marking, particularly where students had an option in what they could research and what their project would consist of (e.g., a written report, a poster, a physical replica model, etc.). Several of the teachers noted that most of their assessments were primarily of "the product" (the lab report, or the paper) as giving evidence of various processes, rather than direct assessment of "the process" (observation of the behaviors and judgment of the skills). Several of the teachers used this product assessment quite frequently.

Performance assessment should be useful to assess important aspects of the curriculum that cannot be readily assessed using paper-and-pencil tests, such as science processes (e.g., observation, controlling variables, hypothesizing, communicating), and social studies thinking and research skills (e.g., interpreting data, drawing inferences). While there was some direct assessment of student behaviors, assessment of products predominates. The teachers were familiar with holistic and analytic approaches to scoring, probably obtained from developments in the assessment of student writing in language arts. Teachers could well use assistance in developing systematic observation tools and other procedures for the direct assessment of process skills (e.g., Stiggins, 1987a). Developments in the area of performance assessment provide some suggestions and guidelines for their application in the classroom: Berk (1986), Stiggins (1987b), and Haynes and Wilson (1979).

**Oral questioning strategies, interviews, and presentations**. This category of assessment methods included oral questioning strategies during instruction for purposes of guiding instruction, and teacher-student interviews and oral presentations by students for purposes of assessing student learning. Although these are quite distinct procedures they were combined to simplify rating of importance by teachers. The average rating was low, and ranked seventh out of the eleven methods. The teachers generally used a considerable amount of oral questioning during actual instruction, as well as during review before a test and after a test or assignment, but there was considerable variation among the teachers. Questioning strategies included open situations where students volunteer to respond and the teacher chooses one or two, and directed situations where the teacher asks a specific individual. It appeared that all students were involved in both situations, although typically there was no formal plan or schedule used by the teacher to ensure the involvement of all students and no systematic records were kept (some anecdotal notes were kept). All four teachers indicated that they consciously attempted to involve all students in the questioning.

One teacher suggested that it would be useful for beginning teachers to develop an in-class questioning plan and maintain a record of both her/his questioning pattern and student responses. The teachers frequently had specific reasons for eliciting the response from a particular student; the reasons varied from one situation to another, ranging from trying to elicit student thinking to prompting student attention. All four teachers identified students for responses based on reasons such as: diagnosing possible student difficulties; motivating students by giving them opportunity to participate and to demonstrate what they know; and attempting to develop critical understanding, to elaborate on what others have said, and to make specific applications of concepts. The teachers considered it an important teaching strategy both as a learning tool for students (e.g., elicit their thinking, reinforce learning) and for monitoring pace and focus of the instruction (e.g., to determine if students are "on the right track").

Two of the teachers indicated that they use some oral presentations by students as part of their assessment program, but this was not frequent. None of the teachers used oral presentations during the observation sessions. One teacher provided oral presentations as an option for students in giving their project report, and encouraged this and other alternatives to the written format (such as video tape instead or written report). Several of the teachers indicated that they have used checklists and other procedures to assess oral presentations, and also have involved students in the process. The teachers varied in their opinion of the value of oral work, with two of them strongly advocating it.

Student interviews were not typically used for assessment purposes by the teachers. One teacher did note that he had used interviews in cases where the student had an obvious difficulty with written language: the teacher described one example where he administered the classroom test orally to the student and the student responded orally.

Certainly oral interviews could be used by teachers to assess students; for example, written tests could be read aloud. However, individual interviews are very time consuming, and may have only limited application in the classroom setting. Oral questioning of students as a group and oral presentations by students are not as impractical, and are important strategies in the classroom. The teachers did not indicate that they had any systematic training in the use of these procedures, and there would appear to be merit in preparing teachers for all three aspects of oral assessment.

**Standardized tests.** Standardized tests (including those initiated by the school, division, and province) were simply not used in the classroom by these teachers. The teachers indicated familiarity with various types of standardized tests (e.g., survey achievement tests, ability tests, individual intelligence tests) and could give some potential uses. However, they had little if any formal training in the use of standardized tests, and were not able to deal with some of the critical concepts such as the interpretation of standardized scores.

The teachers were ambivalent about potential uses of standardized tests in science or social studies. One teacher pointed out that periodically it would be valuable to compare the performance of his students to that of other teachers' students to determine if teaching and expectations deviated. This, of course, would require a comprehensive test directly relevant to the course and program. In the main, the teachers saw standardized tests as not being sensitive to topics in the courses they were teaching, nor to the way in which they were taught. The teachers questioned the validity of using the same test across several teachers unless they had cooperated in the development of it. What the teachers thought would be useful was to review other teachers' assessment materials. They saw provincially developed achievement tests as having value, perhaps more as a source of good test items, but were not supportive about commercial standardized achievement: one teacher commented, "administration of the CTBS [Canadian Tests of Basic Skills] gives the teacher three days of nothing to do". At one time these tests were administered at the junior high level in the district, but the teachers had no recollection of using the results. The teachers typically did not review student cumulative records, where information from standardized testing is commonly recorded, but indicated that at times this information could be useful to help illuminate a problem with a particular student that they may have noted.

**Group assessment methods.** Students working in small groups is a common practice in junior high classrooms (usually groups consist of two or three students), and these teachers were no exception. The two science teachers frequently had students working in groups in the laboratory. Partly this was a function of limited resources, such as equipment, but it was intended also to teach students to cooperate and to solve problems. The social studies teachers grouped students for projects or other types of exercises. All four teachers noted problems with marking group-based assignments, such as identifying clearly who should receive credit for what. They also indicated other problems such as inequitable work effort of group members, student motivation and frustration in certain groups, and organizational/logistical problems (e.g., students absent). The teachers all structured group assignments very carefully and monitored progress on a regular basis: the task and assignments were usually clearly written out, and timelines and other details given. Group structures and purposes, such as those recommended by Johnson and Johnson (1991) were not applied in any systematic way, and groups were seen primarily as a vehicle to make more efficient use of limited resources, to provide for variety in instructional approaches, and to enhance motivation.

Most of the group assignments in science resulted in individual reports which could be marked individually, but this did not obviate the problem of lack of independence. All four teachers marked individual products for some situations (one teacher in particular), whereas for other situations marked a common report. For example, in the laboratory a general problem was set, such as determining when photosynthesis occurred in plants. Students could select one or two partners to work with, and could divide the work load, but each student's report was marked according to a set of criteria. Further, the group members were observed and rated on the laboratory skills, safety practices, and on various other behavioural characteristics. The teachers had concerns with assessment in group settings, and, for example, one teacher stated that he had tried many different ways of using groups, but was not satisfied with any of these and was continually changing. He also had used procedures by which groups monitor themselves, both in terms of behaviour and achievement.

Several teachers noted the value of group work for developing cooperation and cohesiveness, which they saw as being an incentive for student involvement and effort. One teacher used a system of friendly between-group competition to focus and motivate student effort: groups were formed and given tasks such as identifying the important concepts in a passage and answering a number of questions; the group total scores were then compared.

The teachers indicated that they usually allowed students to choose their own partners and form their own groups. This posed potential problems for evaluation since high achieving students tended to choose each other, and group projects that require division of tasks could greatly favour the high achievement group because of the compounding benefits of more able students doing each task (this too was noted by one of the teachers). Group work is an important part of learning as well as social interaction in the classroom. Many researchers have provided evidence that supports the value of cooperative learning, but have also indicated the complexity of the relationships between achievement and group structure, purpose, and function (e.g., Johnson & Johnson, 1991; Slavin, 1983).

It may be possible to assist teachers in providing guidelines that identify what group structures and functions are best, which kinds of objectives and skills should be taught in this manner, and how assessments can be carried out to reduce unfairness as much as possible. As one example, cooperative approaches can be used to bring about learning, particularly of a interpersonal or social nature, but rewards may be based on individual performance of group members (Crooks, 1988).

**Opinions of other teachers.** As a method of assessment of student achievement the opinions of other teachers were not typically sought. All four teachers expressed concern with the possible prejudicing effects of obtaining information on students from other teachers, and indicated that this information should be interpreted with considerable caution. They saw value in discussing students with other teachers familiar with the student, teachers who teach another course to the student or who taught the student in the previous year, in order to understand the student and to get insights as to how to communicate with her/him. The teachers also noted that the student may not respond the same way to other teachers. They indicated that no assessment for grading was based on this kind of information.

The teachers occasionally sought the opinions of other teachers about a student, or accessed the student's previous record, if the student was having difficulties or if there were problems in dealing with the student. The information most often sought was on students' social development, preparedness for certain tasks, or special problems, and this was sought primarily from colleagues teaching at the same grade level who also taught the student. This information was usually informally sought and used. As one teacher noted, it was more to confirm or check information regarding a student, but the purpose was to develop a plan for the student.

**Assessment of reasoning skills.** None of the four teachers used a taxonomy of cognitive levels, or any systematic procedure, to ensure that higher levels of cognitive skills were assessed. This area was rated fifth out of eleven methods in importance by the teachers (of moderate importance). Although formal structures, like test blueprints, were not used, all four teachers attempted to set questions which presumably tapped more than simply knowledge and comprehension; these were often in the form of problems with some novel aspects. One teacher suggested that about 10-15% of assessment should tap reasoning skills, but the analysis of teacher-made tests indicated that probably less than 10% of test-based marks were directed to items at these levels.

The teachers gave some evidence that they could construct questions that assess higher-level thinking skills. They gave examples of questions that were clearly higher-level, such as "how would you heat water in a gravity-free situation like you have in a spacecraft?" (students had learned about conduction, convection, and gravity), and "are 'mystery' chemicals acidic or basic?" The teachers gave "What if. . . ?" and "Which is correct and why?" questions with simulated situations to which students wrote responses, and they asked students to produce statements connecting several facts and notions. Programs in science and social studies prescribe a number of higher-level cognitive skills (e.g., hypothesizing, experimenting, inferring), and the teachers indicated that these skills were important. There remains the concern, however, that the higher level skills were not systematically assessed on tests, and certainly not to the extent that would be desirable. This was the conclusion of Chambers (1982), Stiggins (1986a), and Stiggins, Griswold, and Wikelund (1989).

**Regular homework assignments.** This included assignments on which students worked in class and may be taken home for completion, and those that they were to do outside of regular class hours and usually take home. All four teachers tried to gauge student work so that most of it could be completed during class hours if the student worked consistently. They also recorded the results of most of the assignments, and these were incorporated in student grades, although the actual weight applied to any one assignment would be very little if there were several assignments for a unit. The number and size of assignments and how these were weighted varied considerably from teacher to teacher and from unit to unit for one teacher. The assignments were usually clearly stated and marked almost immediately (e.g., returned to students the next day). Students were often required to make corrections.

The assignments were based directly on the material, in some cases directly on the textbook, and ranged widely in format. What was lacking in the assignments for at least two of the teachers were questions that required more than recall, such as reading the text and selecting the correct information, on the part of the student. The teachers believed this to be an effective way for students to learn the material and teachers to obtain assessment information: it gave students a chance to work hard and do well, and completion of task was considered an important criterion. Two of the teachers preferred to set more of their own assignments, and one preferred longer assignments, such as projects and papers (he also had older students--grade 9 rather than 7 or 8).

**Student peer ratings.** Peer rating of each others' work or performances was rated very low in importance by the teachers. They expressed concern that peer ratings can be harsh and this may be detrimental to students. They also noted that there are problems with fairness: with the ability of students to accurately judge one another, and with students' maturity to be able to ignore popularity. Several of the teachers indicated using rating schemes and other methods to aid students in peer assessment, but used them infrequently. The teachers noted some important benefits, such as focusing students on deciding and learning what is important, students' learning to observe and to judge, and students' learning to help one another. One teacher noted the importance of making ratings anonymous, of everyone opting willingly into the process, and of everyone being prepared for the task. All teachers felt that many junior high students were not mature enough to do peer rating or to benefit much from them.

This is an area where there are few guidelines to assist teachers on when and how to use student ratings, and what aspects students at various grade levels can be expected to understand and rate effectively. There are clear dangers in the process and these must be well understood by the teacher who intends to use peer assessment of any kind. There are ways students can be involved in the assessment process other than by directly rating the work of other students. One such way is to have students respond to each others' written work by writing what they think they understand from it. This may have less of an judgemental tone, and may not have as potentially dangerous effects. It is also a procedure used in the teaching of writing.

**Student self ratings.** The teachers rated this method at much the same low level of importance as they did peer rating. The teachers did not appear to be well versed in how this could be used, although one teacher indicated that he used this approach on occasion. He suggested that students must be involved in setting the criteria, and must understand the categories for the assessment (what is to be considered in the evaluation).

Some of the same concerns of fairness and maturity that were expressed for peer ratings were applied to self ratings. But there is usually not the same level of potential damage involved, such as the possibly devastating nature of being criticized by your friends; self-evaluation is after all a private process. One teacher noted that he found students to be quite honest in their self assessment, and, if anything, overly harsh on themselves. However, as with peer assessment, there is a distinct need for guidelines and sample procedures that teachers can use.

**Strategies for integrating assessment and instruction..** All four teachers used in-class and homework assignments for instruction and assessment. Assignments formed part of the classroom activities to present new material to students as well as to reinforce concepts that were being taught. The assignments were marked and these marks recorded and included in forming student grades. Also, paper and pencil tests contained items which were near replicas of those in the assignments. This can be considered as a way to reinforce learning, although the effects are not clear (e.g., Crooks, 1988).

The teachers systematically provided review prior to a test and followed the test with a review of selected items on the test. One of the teachers indicated that he spent considerable time teaching students how to answer questions, and part of this instruction included students actually identifying what they thought was important to test and creating items for this. The student-created items were used for review. Material that posed problems for a number of students were typically reviewed in some detail. Items on the test that posed difficulties to students were reviewed and discussed in class, usually the next day.

**Dealing with cheating.** Cheating by students was of concern to the teachers, and all were prepared to deal harshly with offenders. They indicated that a student caught cheating on a test or an assignment would have the paper destroyed. Typically, this was made apparent to students early in the school year. The teachers monitored testing (and often rearranged seating to reduce opportunities for students to read others' test papers) and marked the tests themselves, thereby minimizing cheating within the testing setting. There was no evidence of direct copying of anothers' work during the observation period of the case studies.

All four teachers taught the same course to several classes of students, and largely used the same tests for these classes. Students often obtained information from other classes on the test content, so unfairness may have occurred. The teachers indicated that they occasionally used strategies such as rearranging the test items from one class to another, changing some of the questions, using more supply-type questions with longer answers, and administering the same test on the same day to reduce the effects of students communicating test information to one another.

The teachers stressed the importance of "doing one's own work", and demonstrated to students when and how it was acceptable to receive assistance from one another, such as during group work, and when it was not, such as in tests and individual homework assignments. They also indicated to students that "getting help" in doing an assignment was acceptable provided the student learned how to do the question and understood the answer, but that it was not acceptable to copy from another student.

It appeared that this is an important issue with teachers and that they deal with it seriously. One of the teachers indicated resignation to the fact that there will always be some cheating. But the teachers were concerned with providing a setting which minimized the attraction of cheating, at least during tests. One teacher indicated that he spends considerable time in preparing students for tests, and in building their confidence in themselves and pride in their work, so that cheating becomes less than acceptable. He believed that students in this kind of setting are remarkably honest and can be trusted, but also believed that cheating by copying another's work must be made more difficult so that students do not inadvertently cheat or slip into cheating because it is easily done.

The fact that the teachers were experienced was apparent from the observations and interviews. Their attempts to reduce the likelihood of cheating meant that they rarely had to deal with it after the fact.

### 3. Methods Used by Teachers for Assessing Affect

The assessment of affect was perceived by the teachers as being of considerably lesser importance than that of assessing achievement, and although teachers saw the affective domain as important to learning in the classroom it was not commonly assessed in any systematic way. The affective domain contains traits that can be considered of two different orders in terms of classroom activities and instruction. The first consists of characteristics that relate to general deportment in the school and classroom and are relatively independent of the particular curriculum and course being taught. These are based on generally accepted values in our society, and examples are diligence and effort, honesty, integrity, obedience, cooperativeness, tolerance of and respect for others, willingness to learn and to do well, patriotism, and the like. These are what Gullickson (1985) would classify as "citizenship and behavior", and often they fall under the general description used on student reporting systems, "she/he has a good attitude".

The second order of characteristics consists of values, attitudes, and behaviors that could be considered curriculum-related, although many of these attitudes are reflected in goals of more than one subject area. They also are often based on generally accepted societal values. They include such things in science as withholding judgment, careful and objective observing, being swayed by evidence rather than beliefs, sensitivity to the environment, and importance of safety practices, and in social studies, acceptance of alternate views, tolerance of others from differing cultures, willingness to accept information from alternate points of view, and importance of social involvement and social responsibility. The teachers indicated that they often stressed affective goals in their teaching, and used such activities as class discussions to bring them to consciousness, but rarely did the teachers actually assess the associated characteristics of their students.

**Observing individual students**. The teachers frequently observed the behavior of individual students, and in some instances made affective assessments of it. This was rated as the most important method for assessing affect. Observation tended to be done on an informal basis and usually affective behavior was noted only if there was some unusual or disruptive event or if the student repeatedly behaved in a way that was cause for concern. An example of this was that a student would be reprimanded if she or he interfered with the work of others, but this was recorded if the behavior was repetitive. One teacher noted that he did not explicitly assess such traits as scientific attitudes but did occasionally assess such aspects as lab safety, care and cleanliness in the lab, and the use of scientific processes such as systematic observation procedures. Some affective aspects of learning become embedded in the assessment of achievement: for example, typically more marks are awarded for assignments that were completed on time, neatly done, and free of errors.

The teachers did not identify a list of affective traits that they monitor, but often gave students oral feedback on various characteristics. One teacher asked students for their beliefs, attitudes, and concerns regarding both classroom activities and the subject matter. This was used more to understand the students and what particular students may be thinking and feeling, and typically was not treated evaluatively. The provincial curriculum documents identify the kinds of attitudes that are considered important, such as dispositions in science (e.g., withhoding judgement, basing beliefs on evidence) but guidance is needed as to how these are to be brought about in the classroom, what constitutes behavior consonant with the desired attitudes, and how this can be monitored and communicated.

**Observing group interactions.** Although group activities were fairly common in these teachers' classrooms there was little systematic attempt to assess affective aspects of the group interaction. The teacher who used this kind of assessment focused on the listening and participation of students in the groups and also in the class as a whole. Data were gathered informally and usually only sporadically recorded, but he did attempt to assess the level of participation.

As with individual assessment, guidance and support are needed for teachers. Procedures have been developed for systematically observing students in groups and these can be adapted for teacher use in the school context. An example is time sampling of behaviours, and specifying behaviours in advance and using checklists or rating scales.

**Using questionnaires.** Although teachers indicated that questionnaires were quite important for gathering affective data, by rating this category 20 points out of 100, it appears that actual use is infrequent for affective evaluations. One teacher used an anonymous questionnaire to ask students such questions as: "Do you like where you sit?", "What could you do to improve your learning?", and "Do you think this test was fair? does it ask you what we have studied?" The teachers appeared to recognize the difficulties of questionnaires, and noted such problems as: possible student inability to answer the questions, possible bias of questions, and sensitivity of self-disclosure information. It appeared that the teachers valued the comments students make but, in the main, did not actively seek them in some systematic way.

Questionnaires are commonly used in research settings and guidelines for their construction are available. These could be adapted to the classroom, and teachers instructed in their use.

**Using interviews.** Individual interviews of students were seldom used. One of the reasons given was that it takes an inordinate amount of time. The teachers readily identified some important traits that could be assessed in an interview--attitudes, beliefs, and prejudices, and anxiety, lack of attention, and other problems stemming from personal situations. But difficulties noted with interviews, such as need for close rapport with the student, often confidential nature of the information, and amount of time required for student to "open up", worked against their use. Casual conversations with students were commonplace and these may have provided information to the teacher about how the student feels about something. There was little systematic information gathering and recording. Interviewing seemed to be used when a student was perceived to be having problems. Usually this meant problems with achievement in the course and the interview was intended to help the teacher in diagnosis, but the problems could also have been social or personal and the teachers would attempt to determine this.

Effective interviewing of school-age children is a skill that can be taught and should be available to teachers, particularly for obtaining information on affective characteristics.

**Opinions of other teachers regarding affective characteristics.** All four teachers mentioned the possible biasing effects associated with obtaining information about students from other teachers (this was noted also with respect to the assessment of achievement). Although this method of obtaining affective information was rated slightly higher than four others, it still only warranted 7 points out of 100 on average. There was a considerable amount of informal discussion, sharing of information among teachers regarding students' achievement, attitudes, behaviours, and problems. This was common among teachers of different subjects at one grade level but sometimes involved previous teachers of the students.

The primary reason indicated for seeking opinions from other teachers was to determine if there are problems or concerns with students, or ways teachers might more effectively deal with a student. One teacher provided an example: a student responded negatively to any pressure from the teacher, but another teacher indicated that he could successfully motivate the student by coaxing and gentle persuasion. In both schools teachers worked in teams and thereby communicated frequently regarding the behaviour of students. As with the other methods teachers could benefit from more systematic communication of affective characteristics of students.

Other sources of information for affective characteristics. The teachers did not rate highly opinions of other students and parents, and student cumulative records, for obtaining affective assessment information. The teachers typically did not solicit information about the affective characteristics of a student from other students, although they may ask a question like, "Do you know what is troubling Leslie?" This type of information was used, at most, as a starting point for getting further information about a student, and certainly was not sought systematically for the majority of students. The teachers expressed concern about the possible prejudiced, and prejudicing, nature of information from these sources.

The opinions of parents were sought more systematically, and there are procedures in place in most schools for teachers to meet with parents, although they are usually designed to center on reporting student progress. Parents were also encouraged to comment on their children's report cards. Again, though, the affective information obtained from this method was only used when there appeared to be a problem that the student was having (particularly as evidenced in achievement or in classroom behaviour).

The teachers indicated that past student records were typically not used until after they had become familiar with their students, and then only in particular situations where there was some problem arising with a student. One teacher noted that student records could provide hints as to previous behaviour patterns and suggest possible problems a student may be having, but he preferred to seek primarily positive or constructive comments about the student (e.g., "Leslie works better when given a clear task to do"). The teachers indicated that information on affective characteristics, such as endeavor and effort, willingness to work, and cooperativeness/disruptiveness, were part of a student's cumulative record. But all four teachers preferred to obtain affective information on their students directly from interaction with their students rather than from other sources.

## Assessment of Skills by Teachers

None of the four teachers initiated administration of any standardized, published ability tests nor did they use information from them. This nonuse was discussed earlier in the context of standardized tests and achievement, and is well documented in the literature for teachers across North America (e.g., McLean, 1985; Stiggins & Bridgeford, 1985). Typically the teachers did not separate ability from achievement in the course material they were teaching. Ability was construed as the ability to perform well on course material, on tests and other forms of assessment in that course. Several of the teachers emphasized the importance of cognitive skills more general than those directly associated with the content of a particular course: such as careful, interpretive reading skills (of course-related material, usually), systematic reasoning skills, and drawing and graphic skills. These teachers appeared to make an attempt to incorporate assessment of these types of skills in their assessment materials for the course; one example was the presentation to students of a written passage and test questions based on their interpretive reading of the passage. Another was the requirement in one class for students to represent pictorially what they saw through a microscope. All four teachers mentioned the importance of writing skills.

This concern often appeared in the directions to students on assignments and tests, and sometimes was incorporated in the marking. In the main, however, there was little or no direct attempt to measure ability constructs separate from achievement in the course.

It is worth noting that teachers incorporated the notion of constructs, such as general ability in their understanding of students, and often made comments like "I have a very bright class this year" or "This class is weak so I spend a lot more time explaining the concepts to them". Other constructs that commonly appeared to underlie the teachers' perceptions of students and classes were reading ability, writing ability, mathematical ability, and the ability to work independently and with little structure from the teacher

The teachers made little use of formalized ability assessment, and thus there was little impact of this type of assessment on the classroom. There was no direct ability grouping of classrooms in the two schools, although the teachers indicated that they may adjust both their teaching and assessment according to the perceived ability levels of particular classrooms. These ability estimates were based on achievement of the students either in the course that is being taught or in courses from the previous year. The teachers continued to use assessments in the course to confirm or dispute their perceptions, and often commented on one class being more or less able than another based on the assessments. One teacher systematically compared the average performance of classes on tests, in one year and from one year to the next, and used this to motivate students.

Assessment of more specific abilities was apparent, particularly that of writing ability. All four teachers spent time instructing their students on how to write, on the mechanics and structure of writing, and occasionally on what is important to say and how to say it effectively. The teachers based their concerns with students' communication skills, particularly writing ability, partly on the work of students in their classes, but also on the general belief that these are extremely important skills that must be developed in their course as well as in language arts.

Some of the assessment in content-based courses are dependent on communication skills, such as the preparation of a laboratory report (and other skills as well, such as interpreting graphic or pictorial information). It would be useful for teachers to identify the skills that are important to develop in the context of the course and to use systematic procedures for assessing these skills. In part this has been done through the curriculum specification of processes in science and thinking and research skills in social studies, but more support is needed in how these often elusive constructs are to be assessed. There are procedures available from other contexts and subject areas that may be of use, such as the approach to assessing written work suggested by Biggs and Collis (1982), and direct assessment of writing using holistic and analytic procedures (e.g., Huot, 1990). Procedures to assess more general cognitive skills are available in sources such as Norris and Ennis (1989) and Stiggins, Rubel, and Quellmalz (1988).

## 4. Factors and Criteria Used in Deriving Assessment Plans

The relative importance given to nine criteria teachers use for the selection of assessment methods and materials was discussed earlier. The purpose for the assessment and the learning outcomes to be assessed were the primary criteria identified by the teachers, and the assessment of higher-level thinking skills was well down in the ratings. The observations and interviews regarding these and the other criteria are given below.

**Assessment results fit purpose.** This criterion was rated highest on average by the teachers, but was followed closely by the criterion "method matches intended outcomes". The teachers indicated that most assessments were to promote and reinforce learning (particularly assessments during instruction, such as assignments, text questions, papers and projects, and laboratory exercises) and to provide information for grading and reporting. Some note was made of the need to evaluate instruction for modification of teaching. Thus, much of classroom assessment was selected or developed with this focus in mind. The structure of classroom assessment suggests that this is so: the considerable emphasis on assignments (they are part of nearly every day's instruction), and the marking of these for inclusion in the grading process, illustrates this focus.

However, there are more and less efficient methods for accomplishing these purposes. For example, careful sampling of the content domain is a necessary feature of efficient assessment for grading and reporting student progress. This is not as effective for motivating students or for promoting learning. Much more detailed kinds of assessment are needed, which may not reflect accurately the respective importance to the course of the concepts being assessed. The validity of the assessment procedure chosen depends on how the results are to be used. Many assessments serve more than one purpose, and this is often reasonable, but the structure required for one purpose is often not the most efficient for another.

**Assessment methods match intended learning outcomes**. The learning outcomes, the goals and objectives of the teaching, usually provide the content to be assessed--and this is as it should be. Teachers view this as an important criterion for their assessments, and often criticize standardized tests because they do not reflect exactly what was taught, what examples were used and how the topic was developed. There is concern by teachers and by curriculum specialists that the format and method of assessment should be congruent with the outcomes of interest: for example, it is now commonly believed by language arts experts that proper assessment of writing should involve actual writing by the student, and the writing tasks in an assessment of writing should be similar to those involved in the curriculum goals and in the teaching (e.g., Huot, 1990). The analogy could be made to science processes: the assessment of laboratory skills (e.g., controlling variables in an experiment) should require students to perform these skills in an actual laboratory situation. This is what Stiggins is calling for when he advocates applied performance assessment (Stiggins & Bridgeford, 1985; Stiggins, Conklin, & Bridgeford, 1986), and similar to what Wiggins (1989a, 1989b) refers to in authentic assessment. The same analogy can be made for social studies, particularly for what are called "thinking and research skills" in the Manitoba curriculum (such as "gathering and interpreting data" or "inferring and drawing conclusions").

The teachers were aware of the effects of various assessment methods and types of test questions. These are some of the illustrations they gave: setting problem situations with novel features for assessing higher-level thinking, ratings of behaviours for assessing affective traits such as cooperativeness or safety, actual testing in the lab of chemicals by students for assessing lab skills, and items that require students to supply such things as diagrams or graphs for assessing the ability to present information. The analysis of tests that follows suggests that the bulk of test items were of the short-answer and completion variety, that are not conducive to assessing many of the skills and objectives of the program. This was not in keeping with the goals of the programs--a much greater emphasis on higher-level thinking skills and affective and behavioural learning is advocated.

Thus, despite the importance attributed to this criterion for the selection of assessments, and the apparent knowledge teachers have of assessment methods, this may not translate into common practice on tests. In fairness, it should be stated that tests only form a part of the total marking and grading of students (typically 30-40%; see example in Figure 1 above). But this would mean that much of nontest assessment should be at the higher cognitive levels to compensate its absence on tests, and some also should involve affective and behavioural assessment. Discussions with the teachers and observations in their classrooms did not bear this out. Some of the nontest assessment was based on longer-term papers and projects, but there was no assurance that these assignments required higher-level thinking. Some undoubtedly did, and some required such processes as gathering and presenting information, but not all. Many of the assignments given in class had questions very similar to those appearing on tests, and would fall prey to the same criticisms. There was a substantial amount of lab work done in the science classes and these lab assignments did require a variety of processes (such as observing, recording, testing, etc.), but it was difficult to determine the overall amount of this. Students appeared interested in the lab work and certainly took it seriously. It is now believed that it is necessary to support the hands-on approach to teaching science by using assessment techniques that tap these important skills and behaviours, and that can be used directly in the laboratory and other settings, such as outdoors (e.g., Baxter, Shavelson, Goldman, & Pine, 1992; Yager, 1989).

A similar case can be made for social studies. The Manitoba curriculum specifies objectives in the major areas of thinking and research, attitudes and values, and social participation. Most of these objectives cannot be assessed in the typical paper and pencil test made up primarily of short-answer or completion items. For example, the social studies objective "effective participation in school and society" would probably require some form of observation schedule or checklist by which the teacher and students could monitor progress and learning.

**Less important criteria for the selection of assessment methods.** Following the above two criteria in importance, five criteria obtained similar average ratings by the teachers. Two of these related to practical considerations of time: "ease of development" and "ease of scoring" were rated approximately equally by all four teachers, 9 to 10 points. Although these ratings were considerably lower than ratings for the first two criteria, they indicated that time was important. This appears to be reflected in the high numbers of short-answer format items on teacher-made tests--they are easier to construct than selection-type items yet can be marked quickly. Several of the teachers acknowledged this feature of short-answer items. The teachers also indicated that with short-answer formats more material can be assessed in a given amount of time than with long-answer and other more complex formats. Since these teachers did most of the marking themselves, time required to mark student papers can become an important consideration, particularly with large classes (25 or more) and tight reporting schedules. The teachers emphasized the importance of longer-answer item formats (essays, etc.). As one teacher remarked, the extensive marking time is well worth it, and he was more than willing to do it. However, it was not always considered feasible to use this format.

There was considerable variation among the teachers in the weight given to the criterion "origin of assessment", ranging from 1 to 30 points. The amount of use the teachers made of assessment materials that were not developed by themselves also varied greatly. One teacher emphasized the need to obtain material from every source available: "beg, borrow, and steal from every source you can get. . . . through conferences, readings, etc." Another teacher indicated that he rarely used assessment material that he did not develop himself, and to him the criterion was not that important. It was difficult to determine how critical the teachers were of the assessment material that they had available to them, and whether the source of the material had much of a bearing on their decisions. Several of the teachers used assessment questions that accompany the texts that they used in class, and thus, this criterion is important since there is no guarantee that the questions are of high quality (probably they are not of high quality, as was suggested by Stiggins, 1986a).

The criterion "degree of objectivity" was of some concern to the teachers. The teachers emphasized the importance of using more than one method to obtain assessment information, and suggested providing students with different situations in which to display their understanding. Usually systematic marking procedures were specified in advance for papers and projects, and were made known to the students, so that marking was more systematic and probably less dependent on teacher biases or quick impressions. The teachers typically used several tests and assignments for each reporting period, which in the main included some objective-type items. These procedures can enhance fairness of assessment, but do not obviate the problem of reliance on the teacher for the choice of assessment materials. Several of the teachers also emphasized the importance of certain program objectives which can only be assessed using subjective methods, such as those based on observations.

It appears that objectivity is a concern of teachers, but perhaps it should be more so since the choice of assessment materials is dependent on the teacher, and since many of the classroom assessment methods used by teachers require subjective marking. There was reported to be considerable collaboration among teachers, particularly within a subject area, on such aspects as course content, instructional material, and assessments. This may provide some freedom from teacher bias in the choice of assessment materials. But there was no reported collaboration among teachers in marking subjective test items or assignments. There is considerable literature available on marking written material using systematic procedures and multiple markers which could be readily adapted to other subject areas and assessment settings (e.g., Huot, 1990).

The last of the five criteria that received a similar weighting was "applicability to higher order thinking skills" (approximately 9 points). This level of importance was affirmed by what the teachers stated in the interviews, although there was considerable variation among teachers both in the weights and in their comments. Higher-level thinking skills were thought to be important, but, according to one teacher, only a relatively small portion of the assessment should be targeted to them--10% or so. The teachers could differentiate assessment methods as to their effectiveness in assessing at a higher level. They suggested student constructed-response items as being more effective: for example, items that required students to justify or give the basis for their response, and questions that required making sense of some material that was presented. As argued earlier, these types of questions are not abundant on teacher-made tests, and perhaps this criterion should receive greater emphasis than these teachers (and others) afforded it.

**Criteria considered unimportant for the selection of assessment methods**. The teachers considered the criteria "time required to administer" the assessment and "effective control of cheating" of relatively low importance (4-5 points out of 100). These teachers used a substantial amount of classroom time for assessment purposes, but apparently were not concerned much about this time. It would be useful to bring this to the attention of teachers, and for them to determine on an explicit basis as to whether the amount of time used for assessment was reasonable. Some assessment methods are more efficient in terms of student time, and others in terms of teacher time in preparation and in marking.

The problem of cheating was important to these four teachers although there was some variation among them in the weights given this criterion. The teachers identified procedures which reduced the opportunity for cheating, particularly in testing situations, and noted the problem with assignments and projects. There was a realistic view of cheating, suggesting some resignation, and one teacher stated: "I guess, if a person wants to cheat they're going to cheat". However, the teachers attempted to make the work important to the students and thereby elicit honest effort. One teacher was very clear in expressing his view:

> I don't like any kind of cheating, and I tell my students if they ever see the dark side of me it's in that area. . . that really bothers me. . . . If your kids are cheating on you I think that the teacher has failed, because you haven't provided an environment where kids would feel confident in doing something else.

## Quality of Classroom Assessments

The assessment instruments of the four teachers were analyzed to determine technical characteristics and quality, and the levels of cognitive learning which the assessments were likely to tap. Instruments consisted mainly of teacher-made paper and pencil tests (including quizzes), although performance assessment, and oral forms of assessment (including interviews and other types of oral questioning) were also considered. There were only several alternative assessments, and since the procedures for these were not written down they were not included as part of the review that follows. Paper and pencil tests and performance assessment probably comprised upwards of two-thirds of the assessment used for student grading (see Figure 1 for an example of how various assessment devices are included to obtain grades). Standardized tests were not included in this analysis since they were infrequently used by the teachers (as is true generally for teachers at these levels). Self- and peer-evaluations were also infrequently used, and so these were not included for analysis. In-class and homework assignments were given on a near-daily basis, but they were not analyzed since it was nearly impossible to gather the specific information necessary to conduct a systematic and complete analysis: for example, some of the assignments were given orally or written on the board and thus not available without observing classrooms continuously. The importance of daily assignments to evaluation must be noted, however, since they contributed 20 to 50% to student grades.

In total, 35 documents from the four teachers were received for analysis. Of these, 33 were paper and pencil tests that were administered to students during class time, whereas two were project assignments that required student work over several class periods and possibly some work at home. The test documents consisted of 7 quizzes (typically based on one-two weeks of course instruction), 14 chapter tests (several weeks to a month of instruction), and 12 unit tests (several chapters of work). Although the teachers were asked for all of their test materials for a given unit of the course or reporting period, a complete set was not obtained from each (a reporting period was one third of the school year in one school and one quarter in the other). The teachers were asked on several occasions, but all documents were not forthcoming. However, the teachers stated that what they submitted was representative of what they use regularly in their classrooms for the particular course. All four of the teachers provided a general outline of the relative importance of the tests in their student evaluation program and the approximate weights that they attach to tests and quizzes. This typically ranged from about 25% to 50% of the mark for a grading and reporting period, although all four teachers indicated that this varied considerably from unit to unit in the course. The example given in Figure 1 indicates how the teachers typically incorporated tests into grading schemes.

### Paper and Pencil Assessments

The 33 tests were reviewed in terms of structure and analyzed for technical quality as test documents. Second, item formats appearing on the tests were summarized. Third, the items were analyzed to determine the probable cognitive level they required of the student for a correct response. Finally, the test items were analyzed as to their technical quality. This last analysis appears under a separate section.

Much of the analysis depended on subjective judgement of the researcher, so when approximately 15 tests had been analyzed one chapter test that had been analyzed was inserted in the stack of tests yet to be analyzed and analyzed again. The analyses were conducted over a period of several weeks so there should have been little carry over from the first analysis of this test to the second. In fact, this test was not noticed as one that had been analyzed while it was being reanalyzed. The results of the reanalysis were nearly identical with those of the first analysis; the details of this are given for each of the four analyses in the section in which the analysis appears.

**Quality of the test document.** Each document was rated as to its technical qualities, these qualities being identified as questions on 13 criteria (based on Gronlund, 1985; Chambers & Fleming, 1982; and Nitko, 1983). The rating scale used was adapted from Chambers and Fleming (1982): the instrument was rated on each criterion as yes, mostly yes, generally no, and no. The results for the reanalyzed test were identical for all but one of the 13 criteria. In the first analysis the researcher judged that most of the test was arranged from easy to more difficult (rating of mostly yes), but in the reanalysis this criterion was judged as "cannot be determined". This suggests that in the main the judgements were quite stable, at least for one rater. Table 11 gives the number and percentage of the 33 test documents receiving various ratings on each criterion.

Generally the documents were well structured, clear, and easy to read. Items were grouped by format and numbered consecutively either within the group or for the test as a whole. It was difficult to determine if items were ordered according to difficulty, but on a number of tests what appeared to be more difficult items, ones that required longer written responses, preceded short, objective-format ones. Most of the documents were typed, and all but two of the hand-written were legible. Only a few mechanical errors were found, several of which could have been typing errors. The language used was free of obvious bias, and no continual references were made to specific groups, such as referring to a scientist as *he*. Almost all of the documents had titles which indicated the topic being tested, most were paginated, and some had the date and a space for student names.

Problems were apparent in the clarity of directions to students both for the test as whole and for particular groups of items (criteria 6, 7, and 8 on Table 11). Only three of the documents indicated the structure and purpose of the test, the type of items and their marks, and similar information at the beginning of the test. The time allotted for the test was rarely stated. It could be assumed from observing several testing situations that quizzes usually required less than one class period to administer and were often marked in that period, a period being from 30 to 40 minutes. Chapter tests required a full period. It was not clear how much time was given to unit tests which could be expected to take longer. It is possible for teachers to inform students orally at the time of testing how the test is to be marked, but this and other information could appear on the test and reduce possible misunderstanding by students. Also it could assist students in budgeting their time and effort on items in the test. Detailed directions of this nature provide a record of the information with the test itself for future reference by both teachers and students.

A number of test documents did not provide complete directions for each item grouping, and although in some cases this was self evident, students may not remember the peculiarities of a particular teacher and how she or he marks an item. This problem became readily apparent from the differences in directions for similar item types by different teachers, and in one case, between teachers of the same students. Examples of differences in directions are given below, together with the ambiguities that could arise.

**Table 11.** Rating of Test Documents as to Format and Directions

| | Number of Documents (Percentages in parentheses)[a] | | | |
|---|---|---|---|---|
| Criteria | Yes | Mostly yes | Generally no | No |
| **A. Item arrangement** | | | | |
| 1. Items grouped by type/format | 28<br>(100%) | -<br>(0%) | -<br>(0%) | -<br>(0%) |
| 2. Items ordered from easy to more difficult within group and test | 13<br>(52) | 12<br>(48) | -<br>(0) | -<br>(0) |
| 3. Items numbered in sequence throughout group | 29<br>(94) | 2<br>(7) | -<br>(0) | -<br>(0) |
| **B. Test format** | | | | |
| 4. All items within each item-type group have the same format | 29<br>(97) | 1<br>(3) | -<br>(0) | -<br>(0) |
| 5. All parts of a single item-type kept on the same page | 28<br>(90) | 3<br>(10) | -<br>(0) | -<br>(0) |
| **C. Test directions** | | | | |
| 6. Complete, clear, concise directions for the whole test | 3<br>(9) | 3<br>(9) | 27<br>(79) | 1<br>(3) |
| 7. Clear, concise directions for each item-type | 10<br>(32) | 13<br>(42) | 8<br>(26) | -<br>(0) |
| 8. Consistent point values stated for each item and item type | 21<br>(64) | 3<br>(9) | 1<br>(3) | 8<br>(24) |
| 9. Appropriate space provided for student responses | 24<br>(96) | 1<br>(4) | -<br>(0) | -<br>(0) |
| **D. Test production** | | | | |
| 10. Typed or legibly handwritten, neat; clear reproduction | 31<br>(89) | 2<br>(6) | 1<br>(3) | 1<br>(3) |
| 11. Free of mechanical errors: spelling, grammatical, number | 31<br>(89) | 3<br>(9) | 1<br>(3) | -<br>(0) |
| **E. Bias-free** (ethnic, racial, gender, etc.) | | | | |
| 12. Language is universal and non-discriminatory | 35<br>(100) | -<br>(0) | -<br>(0) | -<br>(0) |
| 13. Reflects attitudes which are universal & non-discriminatory | 35<br>(100) | -<br>(0) | -<br>(0) | -<br>(0) |

[a] In all 35 documents were rated, but totals are not 35 for all criteria since some could not be rated.

1. Multiple-choice and true-false items: on one test the number of incorrect responses was subtracted from the number correct. This was clearly indicated on the document, but most other test documents made no mention of this procedure and students were to assume that the number correct would be used, and that it was to their advantage to guess when they were unsure of the correct response.

2. Short answer and restricted response questions: for most short answer items the required answers were sufficiently focused so that students should not have had problems of ambiguity, but in some cases it was unclear as to whether a one word response was acceptable, or whether elaboration of some sort was necessary. Students were guided somewhat by the space allotted and the marks awarded, but the space provided is sometimes a poor indicator of what is wanted and on 8 tests out of 33 the marks were not stated.

These findings are comparable to those obtained by Chambers (1982) for junior high social studies tests. She reported that items were grouped according to format and that "the vast majority of tests [19 out of 23] were typed or clearly handwritten, neat in appearance, and free of spelling and grammatical errors. . . . All tests included language and attitudes which are universal and non-discriminatory" (p. 24). She also found that over half of the tests had problems with directions and reported that "only two of the tests included point values for test items" (p. 24). Clear, concise directions to students, both for the test as a whole and for item groups and items, is an area where improvement is necessary. Information to the student on the test as a whole should include:

- title, indicating topic of the test (e.g., course name, topic/unit, teacher name, date);
- the focus, purpose, and status of the test (e.g., its importance, whether it counts for grades, how results will be used, etc.);
- what the test contains (e.g., sections, item types, etc.);
- length of the test, amount of time available;
- how the test is to be marked (e.g., how many marks allocated to various topics or sections and item groups); and
- a place for student information (name, class, date, etc.).

Specific directions should precede each item group on the test, indicating:

- what the questions are,
- how responses are to be given,
- length and structure of responses,
- marks allocated to items and items groups, and
- how responses will be marked.

Obviously these represent guidelines, and would not be necessary on every quiz. When students become familiar with the procedures used by their teachers some directions would not be as important as others. But at the junior high school level most of this information could be of potential use to the student and the teacher both during test administration and for follow-up review. At this level, students usually have different teachers for different subject areas, and may encounter as many as five or six teachers on a regular basis. Certainly the information should be present on tests that feature prominently in consequential decisions, such as chapter tests, unit tests, and midterms and final examinations. Admittedly such complete directions could make the test document appear much more formal and thereby intimidating to students. This concern should be balanced against the need for clarity and understanding of what is expected. Much of the information and directions could be presented concisely. With the availability of word processing equipment the directions could become fairly standardized and automatic with little effort on the part of the teacher (most of the 33 test documents analyzed were typewritten, and similar directions could become a near automatic component of every test and assignment).

**Item formats included in the tests.** The 35 documents were described according to the item types contained and the proportion of marks accruing to each type. The description for the test which was reanalyzed was identical for both analyses, which was to be expected since relatively little judgement was involved. Table 12 presents the results of the description. The number of marks awarded for the items gave a better picture of the relative importance the teachers attributed to various item types, but this was at best a rough indicator since tests differ in importance and are weighted differently by teachers in the overall evaluation scheme. Without this weighting, which was not always available from the teachers, it is impossible to obtain a definitive picture, although tests that were more highly weighted, such as unit tests, were typically longer and contained not only more items but more variety in item types. The actual test documents usually contained space for student responses, and a crude indicator of test length is given by the number of pages. The average number of 8.5 by 11 inch pages of a test document was: unit tests--2.5 pages, chapter tests--1.7 pages, and quizzes--1.5 pages (based on 12, 14, and 7 documents of each type, respectively).

From Table 12 it is readily apparent that the teachers favoured various forms of short-answer items on tests. Approximately 46% of test marks were awarded to items of this general format. There were more completion items (164) than short-answer questions (120) but the completion items were awarded fewer marks, 17% as compared to 22%. It is notable too that over twice as many test documents included short-answer or completion items than any choice- or selection-type formats, but these teachers also used a substantial number of selection-type items (30% of marks): true-false, multiple-choice, and matching formats received 10%, 8%, and 9% respectively of marks. Just over 10% of test items were context-dependent. They involved interpreting textual information, identifying and labeling diagrams and maps, and interpreting weather maps. There was considerable variation in the amount of material that was to be interpreted, ranging from several pages of textual material to paragraphs and rather simple diagrams.

Based on junior high social studies tests, Chambers (1982, p. 24) reported that, of 798 items analyzed, 38% were classified as matching, 22% short answer and completion, 19% true false, 15% multiple choice, and 3% essay. These results differed somewhat from those of the present study: the corresponding figures were 9%, 39%, 10%, 8%, and 14% for the five categories respectively. These apparent discrepancies are reduced somewhat if proportions of numbers of items are considered rather than proportions of marks awarded: for the five categories the proportions of items were 11%, 35%, 13%, 9%, and 5% respectively. There still remain considerable differences in reported use of matching and short-answer/completion items. The Chambers figure, that as much as 38% of teachers' test items in social studies were matching items, appears high. It is possible that matching exercises are used more by social studies teachers than by science teachers, but this was not born out in the present study. The two social studies teachers had less than 15% matching items. Teachers in the present study clearly favoured questions that required constructed responses, including short answer and restricted essay; this was confirmed in the interviews.

It is difficult to relate the present results to those reported by Gullickson (1982, 1985) and Stiggins and Bridgeford (1985) since their studies involved teacher ratings of the importance of various techniques and did not ask them to indicate the relative amount of this use. Gullickson (1985) reported that grade 7 teachers rated objective teacher-made tests quite highly as to their role in student evaluation, 2.5 on a 3-point scale, relative to essay tests (1.4), oral quizzes (1.3), and standardized tests (1.3). The results of the present study would support these data: objective item formats are greatly preferred--choice and short answer formats combined make up over 85% of test marks. The Gullickson study also revealed that student papers and notebooks (presumably including daily work assignments) were considered important, rated 2.3, and also classroom discussion (2.0). This supports the information of the present study that indicates that tests typically form approximately one third of student marks for grading, and that other assignments, papers, and projects are quite important and are combined in some way to make up most of the remaining portion of student grades.

**Table 12.** Types of Assessment Methods Used by Two Science and Two Social Studies Teachers

| | Number of | | Percentage | Number |
|---|---|---|---|---|
| Paper and Pencil Assessments | Items | Marks | of Marks | of tests |
| **Choice Formats** (selection-type) | | | | |
| •True-false (and T-F correction, yes-no, fact-correct inference-incorrect inf., etc.) | 102 | 102 | 10.3% | 9 |
| •Multiple-choice | 76 | 76 | 7.6 | 9 |
| •Matching | 91 | 91 | 9.1 | 8 |
| •Other alternate-response (e.g., key response) | 46 | 30 | 3.0 | 4 |
| Totals for Choice Formats | 315 | 299 | 30.1 | |
| **Short Answer** (supply-type: word, phrase, statement, number, symbol, etc.) | | | | |
| •Short answer | 120 | 222 | 22.3% | 19 |
| •Completion/fill-in-the-blank | 164 | 166 | 16.7 | 21 |
| •Identification/association (e.g., give the symbol for each element in a list) | 44 | 35 | 3.5 | 6 |
| •Other types (e.g., Cloze, interlinear exercise) | 25 | 31 | 3.1 | 4 |
| Totals for Short Answer | 353 | 454 | 45.7 | |
| **Essay Formats** (supply-type: written, diagramatic, graphic, or other forms) | | | | |
| •Restricted response (limits substantive content and form of the response) | 43 | 138 | 13.9% | 16 |
| •Extended response | - | - | - | - |
| Totals for Essay Formats | 43 | 138 | 13.9 | |
| **Context-Dependent Formats** | | | | |
| •Objective interpretive exercise | 8 | 9 | 0.9% | 2 |
| •Short answer | 88 | 88 | 8.9 | 6 |
| •Essay | 6 | 6 | 0.6 | 1 |
| Totals for Context-Dependent Formats | 102 | 103 | 10.4 | |
| **Total Paper and Pencil Test Items** | **813** | **994** | **100%** | |

What is notable as well from Table 12 is that extended response essays did not typically appear on tests by the three teachers in science and social studies. Essays are assigned but not usually in test settings. Gullickson (1982) argued that the preponderance of objective item formats on teacher-made tests tends to emphasize lower-level thinking skills. Certainly context-dependent formats are relatively infrequently used on tests, in the present study only some 10% of marks were based on this type.

Stiggins and Bridgeford (1985) report that 68% of grade 8 teachers were "comfortable" in their use of teacher-made objective tests (this was also reported for grade 2, 5, and 11 science teachers). Also, 62% of grade 8 teachers were "comfortable" in their use of structured performance assessment. The present results are in keeping with this finding, but it is difficult to relate "comfortable use" with relative importance and amount of use of various assessment techniques.

The results of a survey of teachers by Webster (1987) appear to give a similar picture at least with respect to teachers' belief in teacher-made tests: across all grade levels 50% of teachers rated teacher-made tests as "satisfactory" (3 on a 4-point scale) for assessing student learning, and 48% rated them as "very good" (4 on the scale) (p. 19). The teachers rated "observation of student work/assignments" even higher: 26% rated it 3 on the same scale, and 70% rated it 4. But these same teachers rated "observation of student performance" highest: 29% gave it 3, and 70% gave it 4! It is likely that teachers interpreted "observation of student performance" differently from the way that Stiggins defines applied performance assessment, thus making these results difficult to compare as well. However, teachers in the Webster sample indicated the greatest familiarity with teacher-made tests (90% rated it 5 on a 5-point scale) when compared to other techniques such as observation (50% rated it 5), oral interviews (31% a 5), checklists (23% a 5), and rating scales (12% a 5). It is clear that teachers do value their own tests and feel that they are most familiar with them, but perhaps they value daily classroom work and assignments even more highly.

Webster (1987) reported that teachers made the greatest use of "oral interview" and "work sample", and these are followed in order of use by "observation (anecdotal)" and "teacher made tests of achievement". These teachers represented all grade levels, but there only was a significant difference across grade levels for teacher-made tests of achievement--teachers of grades K to 6 used fewer tests. This serves to emphasize the importance teachers attribute to sources of evaluative data other than tests, particularly to assignments and observations. Webster (1987) also presented the reported frequency of use by teachers of various item types: "completion (fill in the blank) and short answers are used on most tests by nearly 50% of the teachers" (p. 43), and essay items are far less popular. This pattern of use was borne out in the present study, where the ratio of marks for items types was 54% for short answer of all formats, compared to just over 14% for restricted essay.

Stiggins, Griswold, and Wikelund (1989) reported in their review of 149 assessment documents and over 4000 exercises from all grade levels that "34% were selection-type items, 54% were fill-ins, 10% were essay, and 2% required some other type of product as response" (p. 237). Although there was some variation across subject areas, which included science and social studies, the results are comparable to those obtained in this study. R. J. Wilson (1990) obtained similar results as well, although variations across grades and subject were also obscured. If performance assessments are not included, his figures are 22% selection type, 56% completion and short answer, and 22% essay. The results of Bateson (1990) were not reported in a way that allows comparison with the present results.

**Cognitive levels required for responding to the test items**. The test items were analyzed to estimate the level of cognitive functioning required by the student to respond correctly to the item. Bloom's taxonomic levels formed the basis for the categories (Bloom, Hastings, & Madaus, 1971). The higher four levels were collapsed into one category since it was extremely difficult to distinguish items according to these levels. However, it was useful to distinguish items at the higher levels from those at the knowledge or comprehension levels. Furthermore, it was expected that relatively few items would be categorized above the knowledge or comprehension levels (e.g., Stiggins, 1987a). The three levels were identified as:

1. Knowledge (K): items that can be answered readily by directly recalling learned material. Examples are items such as these, "which of the following is the correct term for. . ?", "give the definition of. . .", or "what are the two types of. . . ?"

2. Comprehension (C): items that require some interpretation of the learned material in the response, such as these: "give an example from your home of energy use", or "which of these characteristics makes carnivores well suited for hunting ?"

3. Higher level (A): items that require students to use or apply their knowledge to interpret material which is in some way different from that which was studied in class. This can be done by providing students with material and asking context-dependent questions. These items cannot be answered by simply selecting the answer from the material or from memorized facts and definitions. Examples include: "assume that you saw an animal in a tree with characteristics. . . , what is it most likely to eat?", and possibly followed by "would you expect this animal to hibernate? why or why not.", or "if you could no longer transport goods such as. . . by rail on the prairies which of these is the most likely alternate transportation?"

It was difficult to determine the level of cognitive functioning students used in responding to an item. It is an inferential task on the part of the researcher. In some cases it is impossible to know what exactly has been taught, what examples have been used, and what students know from other sources. Thus, it is expected that there would be some error. In the actual analysis judging many of the items was quite straight forward, particularly since many required the definitions of terms or the terms for given definitions and descriptions (in the main, knowledge-level items). This was apparent for the test that was reanalyzed, which contained a number of items that involved terms and definitions. For the reanalyzed test, only one difference was found: the first analysis judged the test to contain 17 marks for knowledge items and 8 for comprehension items, whereas the second analysis yielded 16 and 9 marks respectively for the two categories.

In total, of the 994 marks awarded on the tests, 576 were for items classified as knowledge (58%), 321 as comprehension (32%), and 99 as higher-level (10%). These findings are reported in Table 13. They clearly support previous researchers' contentions that teacher-made tests typically assess learning at the lower cognitive levels of Bloom's taxonomy. A relatively small percentage of marks on tests were given to responses that required more than knowledge (essentially memorized material) or comprehension (usually involves only limited or literal interpretation of the material used in learning). Fleming and Chambers (1983) reported that as many as 94% of items on tests from junior high schools were judged to be at the knowledge level. Based on his observations at the high school level, Haertel (1986) stated "even items that appeared to call for analysis or supported argumentation proved in fact to require no more than reproduction of what had been said in class" (p. 16). Stiggins (1986a) came to a similar conclusion from observations at the elementary level. More recently, Stiggins, Griswold, and Wikelund (1989) reported that for grade 7 and 8 teachers of science, social studies, and language arts 68% of items were at the recall level, according to their taxonomy of cognitive levels. The recall level includes Bloom's knowledge and comprehension levels (see Stiggins, Rubel, & Quellmalz, 1988, for the cognitive categories used). The proportion of items in the other levels were 10% analysis, 2% comparison, 18% inference, and 2% evaluation.

Even if some errors were made in classifying test items in these studies, the figures bear clear testament to the view that teacher-made testing does not approach what is desired by curriculum specialists (what teachers themselves often state as important), and what is recommended by measurement specialists: that a substantial part of testing should require more than repetition of what is said in class or contained in textual materials.

**Table 13.** Marks Awarded for Item Types at Various Cognitive Levels

| | Number of Marks | | | Percent of Marks | | |
|---|---|---|---|---|---|---|
| Paper and Pencil Assessment Formats | K[a] | C[b] | A[c] | K[a] | C[b] | A[c] |
| **Choice Formats** (selection-type) | | | | | | |
| •True-false (and T-F correction, yes-no, fact-opinion, correct inference-incorrect, etc.) | 80 | 18 | 4 | 8.0% | 1.8% | 0.4% |
| •Multiple-choice | 61 | 14 | 1 | 6.1 | 1.4 | 0.1 |
| •Matching | 82 | 9 | - | 8.2 | 0.9 | .0 |
| •Other alternate-response (e.g., key response) | 6 | 24 | 1 | 0.6 | 2.4 | 0.1 |
| Totals for Choice Formats | 229 | 65 | 6 | 23.0 | 6.5 | 0.6 |
| **Short Answer** (supply-type: word, phrase, statement, number, symbol, etc.) | | | | | | |
| •Short answer | 105 | 107 | 6 | 10.5 | 10.7 | 0.6 |
| •Completion/fill-in-the-blank | 151 | 11 | 5 | 15.2 | 1.1 | 0.5 |
| •Identification/association (e.g., give the symbol for each element in a list) | 27 | - | - | 2.7 | .0 | .0 |
| •Other types (e.g., Cloze, interlinear exercise) | 12 | 13 | - | 1.2 | 1.3 | .0 |
| Totals for Short Answer | 295 | 131 | 11 | 29.6 | 13.2 | 1.1 |
| **Essay Formats** (supply-type: written, diagramatic, graphic, or other forms) | | | | | | |
| •Restricted response (limits substantive content and form of the response) | 7 | 83 | 52 | 0.7 | 8.3 | 5.2 |
| •Extended response | - | - | - | - | - | - |
| Totals for Essay Formats | 7 | 83 | 52 | 0.7 | 8.3 | 5.2 |
| **Context-Dependent Formats** | | | | | | |
| •Objective interpretive exercise | 3 | 2 | 4 | 0.3 | 0.2 | 0.4 |
| •Short answer | 42 | 38 | 22 | 4.2 | 3.8 | 2.2 |
| •Essay | - | 2 | 4 | .0 | 0.2 | 0.4 |
| Totals for Context-Dependent Formats | 45 | 42 | 30 | 4.5 | 4.2 | 3.0 |
| **Total Paper and Pencil Test Items** | **576** | **321** | **99** | **57.8** | **32.2** | **9.9** |

Note: The cognitive level was difficult to determine for some items, particularly those with constructed responses: how these were marked was unclear. Marks awarded to the item were split between the two cognitive levels which the item most likely required.

[a] Items rated as requiring Knowledge-level thinking, based on Bloom's taxonomy.

[b] Items rated at the Comprehension level of Bloom's taxonomy.

[c] Items rated at the Application level or higher of Bloom's taxonomy.

It is a common contention that certain item types, such as true-false, do not lend themselves readily to testing at the higher cognitive levels (e.g., Gronlund, 1985; Gullickson, 1985; Nitko, 1983), although this view is disputed by some (e.g., Frisbie & Becker, 1991). For the most part objective items types, particularly true-false, matching, short answer, and completion formats, are better suited to assess knowledge and comprehension than the higher cognitive levels. This is not strictly true, and multiple-choice items in particular can be used to assess at the higher levels, especially when placed in a context-dependent situation (e.g., Gronlund & Linn, 1990; Haladyna, 1992). But, as an example, matching exercises frequently involve matching terms with definitions, with examples, or with descriptions of the terms, or involve matching locations on a map with names. These are probably knowledge- and comprehension-level tasks for most students. In the present study, 6 of the 8 matching exercises, accounting for 85 marks of the 91 in total, were of this structure and involved term-definition/description matching tasks.

The proportion of marks allocated for each item type was categorized according to the anticipated cognitive level of the items (see Table 13). Items of the objective variety, both choice and short-answer formats, were almost without exception judged to be at the knowledge and comprehension levels: items of this type accounted for approximately 75% of the total number of marks for all tests, yet less than 2% of the marks for these items were for higher cognitive level items. The essay format items were categorized primarily as comprehension and higher level, and the context-dependent items were relatively evenly split across the three cognitive levels. The analysis of the essay formats should be viewed with caution since errors in judgements more likely occurred here, although it is believed that any errors that occurred would tend to inflate the figure for the higher level category. It appears that Gullickson's (1985) statement is supported and that the objective item types do not tend to require higher level thinking, at least not when in the hands of teachers. And, as Stiggins (1987a) stated,

> Measures of recall of facts and information dominate paper and pencil tests and oral questions. Therefore, while instructional objectives and even the instruction activities may aim at thinking skills, assessments often fail to match the intent. Since students look to tests to understand the teacher's expectations, they see the priority placed on memorizing and respond accordingly. The quality of assessments of higher order thinking skills is very low. (pp. 5, 6)

A comment is necessary regarding other assessments that teachers carry out, the 50-70% of student evaluation which is not based on tests. One could hope that this portion of classroom assessment requires more higher level thinking on the part of students. Observation of a number of classroom instruction periods for each of the four teachers suggests that some of the assessment may very well be at a higher level. All four teachers required students to produce larger or longer-term assignments, projects or papers usually, and often students carried out research to complete these assignments. The research activities varied widely, both from teacher to teacher and from assignment to assignment: some required library research, some actual field observations. Unfortunately, it was not possible to obtain from the teachers clear, precise guidelines as to what exactly was required of students and how these assignments were marked. Several student submissions were reviewed, and it was apparent that students often expended considerable effort on the tasks, and used a wide variety of skills in their work: searching skills, organizing and presenting skills, and even ingenuity and creativity! It is informative to note that all four teachers emphasized the importance of structure in the project report, of appearance and neatness, and of sentence mechanics (grammar, spelling, etc.). In some cases these structural and mechanical aspects formed part of the mark awarded. It appears that some of the important process skills were being assessed by way of larger assignments and projects, and that students were being required to think a little more independently. However, there is concern that this was not done systematically enough from one section or unit of a course to another and from one teacher to another.

The report by Stiggins, Griswold, and Wikelund (1989) suggested that higher level cognitive skills were not being assessed systematically during oral questioning in class, although there were slightly more higher level oral questions than paper and pencil ones. Teachers in the case studies asked a much greater range of oral questions: for example, they asked questions of simple facts, but also asked students to identify what was important in a piece of text, to extend on the responses of others to questions, and to give examples of principles from their own experiences. This agrees with the observations of Stiggins (1986): elementary school teachers asked oral questions at all levels of Bloom's taxonomy during instruction, but their tests consisted primarily of items at the knowledge level (requiring recall). Further, the smaller, everyday kinds of assignments given by teachers typically did not fare well in terms of requiring higher-level thinking on the part of students: few required students to do more than recall information or to locate and reproduce appropriate information from textual materials. Observation of classroom activities indicated that many of the questions and assignments accompanying textbooks were used by some teachers, and these questions were not much different from those that appear on teacher-made tests. This was true for both social studies and science.

Some of the work done by teachers in their classrooms was really quite heartening in this regard. Teachers in science required hands-on activities of students: students used laboratory equipment, conducted experiments and obtained observations, analyzed their own data, and so on. Processes and skills were considered important: for example, careful observation and recording, identification and testing of their own samples, safety, and working cooperatively. The assessment of these skills often was in the work product, the lab report or the accompanying assignment, but teachers also observed actual behaviour for such things as safe practices. There appears, however, to be a need for developing more systematic procedures to assess skills of this nature, for applied performance assessment.

## Technical Quality of the Test Items

General criteria have been established for proper construction of most types of test items by such writers as Gronlund (1985), Mehrens and Lehmann (1984), and Nitko (1983). These criteria formed the basis for rating the technical quality of the test items on all of the test documents. The items were categorized as to type or format using the same groups that formed the analysis in the previous section. The items that formed a group on a particular test document were judged as a whole for each of the criteria using a 4-point scale adapted from Chambers and Fleming (1982):

1 = Nearly all items (more than 85% of the items of that type),
2 = Most of the items (50-85% of items),
3 = Some of the items (15-49% of items), and
4 = Few of the items (less than 15% of items).

The number of test documents that achieved various ratings gave an indication of the quality of the items on the whole, and identified which item types were better constructed and which may have more technical problems. The summary ratings for particular criteria gave indications of what particular problems were occurring in the construction of these types of items. Not all tests containing a particular item type could be rated effectively on each of the criteria, since there may not have been enough items. The data are summarized in Appendix C. Each item type is discussed below.

**Choice format items**. Four types of selection-type items were identified: true-false and related two-choice items, multiple-choice items with three or more options, matching exercises, and other alternate-response items such as key-response items (a key of three or more responses is provided for students to choose from for a number of statements). An analysis was possible for each item type since a number of tests contained each type: 9, 9, 8, and 4 tests for the four types respectively (102, 76, 91, and 46 items respectively).

The true-false item format was considered as being less than appropriate for many of the items in seven of the nine tests. This item format does not readily allow for understanding of principles that are generally true and applicable in most situations, but that require exceptions and specific caveats. An example of a scientific principle that is generally true, but that requires a context for it to be certifiably so, is "hot air is lighter than cold air". These kinds of principles are common in the areas of science and social studies. (Multiple-choice formats are often more effective, allowing for situations with a best response but not an unequivocal truth.) The items were determined to be clearly and simply stated and most could be judged true or false. However, six of the tests contained items whose truth was judged on the basis of aspects not central to the concept expressed in the statement (these are often given the label trick statements), and there were several cases of specific determiners, such as "all" or "none", which gave the statements away. There were few problems with mechanical aspects: true and false statements were similar in length, approximately equal in number, and in no readily discernible pattern.

As with the true-false items, a number of multiple-choice items were not considered the best format for testing the content involved. Items that simply required recognition of terms for definitions, or vice versa, can often be assessed using more efficient formats (e.g., matching, short answer). Often the problem of poor distracters arose in conjunction with the inappropriateness of the item type for terms and definitions. Three or more reasonable foils are difficult to obtain in many instances. The item stems appeared to be clearly stated, free of irrelevancies, and with few mechanical problems. The choices did not fare as well as the stems with many items having poor, implausible, even trivial distracters, some choices containing irrelevancies, and in three of the tests the answers for some items were not clearly best. This may be due in part to using multiple choice items in cases where they may not be most appropriate; writing good multiple choice items is difficult and perhaps it is best to reserve them for situations for which they work best, such as where they are context-dependent on a piece of material (usually they would be at the comprehension level or higher). Again, the more mechanical aspects of item construction did not appear to pose problems in the response choices. They tended to be grammatically consistent, free of verbal clues, and similar in length and structure.

The matching exercises received the lowest ratings of item types of a choice format. Many of the exercises were not considered appropriate for the concepts being tested, which were primarily terms and their definitions or descriptions. Although matching exercises can be used effectively for this kind of task, the definitions must be brief and the premise and response lists should be homogeneous with respect to some meaningful criteria. Homogeneity does not mean simply that all responses are definitions, for example. The choice of terms to be included in the list was by no means homogeneous in many cases. Technical terms were often mixed with names of locations or with historical events. Coupled with this problem was the fact that few exercises gave the cognitive basis for matching, for example, instructing students to "match the geographic term with its definition (or with a local example of it)". This was left to be assumed by the student.

For several of the exercises this was readily apparent, but for some it would have helped the student to focus on what was wanted. Matching exercises also often have considerable variation in the mechanics of how students are to respond; this proved to be the case here with each teacher having a particular (peculiar?) way of doing it. Students could do with more direction on how to respond in the matching exercise, such as where to place their responses, and whether a response could be used more than once. Other practical problems arose: many of the lists were simply too long (as many as 15 and 20 entries in each list), and often the longer statements served as the responses, so more rereading was required of students.

There were several key response exercises. These required students to judge a set of items using three or four categories. In the main these exercises were well constructed with few contextual or mechanical problems. Interestingly, many of the items were at a comprehension level suggesting that this format has merit. An example of an effective exercise of this type would be to provide students with a number of everyday examples of phenomena, these are the premises, and have students interpret them or judge them according to some principle they had been studying.

The conclusions reported by Chambers (1982), based on judgements of test items in junior high social studies, differ somewhat from those above. Although she used a more limited list of criteria, some comparisons are possible. She reported that by far the bulk of items were at the knowledge level (upwards of 90%). She concluded, however, that structure and format were not as much of a problem..

**Short-answer formats**. Short-answer test items require that students supply a brief response, which usually is in written form, a word, phrase, or short statement, but also could be a number, symbol, or limited diagram. A question that asks for the geographic definition of the term "island" would be considered short answer, provided no further description or any examples are requested. Four types of items were included under this general rubric: short answer questions, completion or fill-in-the-blank statements, identification and association items (such as writing the chemical symbol for each element in a list), and other related formats (such as completing the missing elements in a table, or cloze exercises). There were 19, 21, 6, and 4 tests containing each of these types respectively. However, since the "other related format" items operated much like short answer items, and often were interspersed with them in the test documents, these questions were rated in the same group making a total of 22 tests containing short answer or related format items. The identification items were very much like completion items so were rated in the same group making a total number of 23 tests containing completion or identification items.

The short answer item format was considered not the most appropriate format for some items in 13 of the 22 tests. Because many of the items asked for definitions and for terms, they could have been framed in choice formats. This would also serve to reduce some of the ambiguity in possible answers. The questions were stated clearly, for the most part, and free of language or grammatical clues, and most could be answered in a short phrase or statement. In the few cases where numerical answers were requested several items did not identify the necessary precision. Very few items required anything but a verbal or numerical response, although several items did ask for brief diagrams. It was impossible to determine if relevant content only was marked by the teachers.

The most frequently appearing item was the completion type: 164 items, or 20% of all items. Several problems were noted for many of these items. As was the case with short answer items, common among completion items were items that required students to supply the term for a definition or the definition for a given term. But many of the completion items were less clear in directing students in their responses. It was felt that many of the items dealt with relatively trivial information: for example, names of detailed parts of apparatus and technical terms. This is not to deny the importance of terminology, but there was little context for most short answer and completion items, and rarely was the student asked to do anything with the term except to supply it in response to a definition or description. The language used in the statements was determined to be straight forward and at an appropriate reading level, and generally considered to be free of technical jargon or statements extracted directly from a detailed passage in a textbook. The blanks in the statements typically did not interfere with understanding the sentence although they did not always appear near the end of the statement.

**Essay formats.** Items were called essay questions if they required a response that was more extensive than a short phrase or statement. Usually this means several sentences at the junior high school level but the response could very well be a diagram, an algebraic problem solution, or an outline map. In comparison with the short answer example of a definition, an essay item might ask the student to "compare the geographic features of a peninsula with those of an island". The form of the student response did not determine whether it was an essay item or not. The distinction was based on whether the response was constructed by the student and not selected from a list, and whether the construction was sufficiently greater than a one word or one statement response to distinguish it from a short-answer item. Two types of essays have typically been identified (e.g., Mehrens & Lehmann, 1984; Nitko, 1983). A restricted response essay restricts the form and content of the student response: an example is "indicate several ways in which the life of a squire was different from that of a peasant". An extended response essay is more open-ended and gives the student greater freedom in choosing the form, organizational structure, and content of the response. Extended essay items did not appear on any of the tests, although the social studies teachers in particular emphasized that their students are required to write this kind of essay sometime during the school year but not in a testing setting.

The restricted response essays on the tests appeared generally appropriate for the tasks. However, teachers rarely gave a clear indication of the expected form, length, structure, and content of the response, and how marks would be awarded. Some of the items did not appear to require more than simply repeating what was taught in class (as Haertel, 1986, noted as well).

**Context-dependent formats.** The context-dependent item formats were combined for the rating process. Although there was some loss in the precision of the rating of the quality of the item itself, the more important task was to determine if the material was appropriate and whether the items were effective relative to this material. In total 10 tests contained items which required students to make sense of the information which accompanied the items. The information consisted of such things as several pages of text on a scientific topic, several paragraphs of historical information, or diagrams.

Most of the context-dependent items were short answer in format although there were several multiple choice and restricted essay items as well. The item types were rated as not being most appropriate in 6 of the tests primarily because some of the items tapped factual material: for example, asking students to obtain a specific detail from the material. Since context-dependent items are particularly well suited to assess at the higher cognitive levels (Haladyna, 1992), this was considered less than ideal. The items were judged as being efficient for the materials and purposes but the overall evaluation of the quality of the items themselves found some items with problems, these usually being lack of clarity in the short answer questions.

The material was judged to be relevant to the goals of the program and appropriate to the level of the students, although some of the materials could not be judged. Many of the materials contained nonverbal information, usually diagrams and maps, but in over half of the tests the information was not considered to be novel (the material had appeared in very similar form on previous student assignments or in class). Some of the material was brief, some was quite lengthy and difficult to comprehend. In almost all cases the items required students to understand, or at least read, the material to answer the questions.

Concluding comments. The overall quality of test items varied considerably. Many of the items were of high quality both in terms of focusing students' responses and the mechanics of construction. The choice format items were generally quite well constructed with few mechanical errors (e.g., language, structure) but often were judged as being less than appropriate for the task. For example, multiple choice items lend themselves well to asking students to make judgements between options of varying likelihood, something that is common in both science and social studies, but they were often used simply to test knowledge of facts and definitions. Some of the true-false items posed problems since their veracity was judged upon a feature of the statement which was not central to the concept being assessed. Of the choice formats the matching exercises were the most problematic: directions were not complete, material in the lists often was not homogeneous with respect to a meaningful category, the lists were often too long, and many of the exercises were not the best way of assessing the particular concepts.

The short answer formats were usually clearly expressed but in some cases could have more than one correct answer. These formats are useful to assess comprehension of material, but few of the items asked students to give examples of concepts or to make any kind of inferences. In some cases the completion items bordered on the trivial.

Some of the restricted essay questions attempted to assess higher level outcomes. However, they often suffered from insufficient directions to the student as to the structure of the response and the required content. Rarely were the marking schemes identified with the item: some of the items could have benefited from statements such as, "3 marks will be for the way you relate x to. . .".

The material for the context-dependent items was in the main effective, but more extensive use of this approach is desirable. The material could be more varied and novel, provided students are informed of this kind of testing and given some experience with it. The items should be based on the material, and should not be answerable on the basis of common knowledge or testwiseness (e.g., Carter, 1986; Rogers, 1991). Finally, context-dependent formats are well suited to assessing higher-level outcomes, and should be used in this way wherever possible. Simple repeating of imbedded information is not an efficient use of the material or the approach.

## Performance Assessment

Applied performance assessment refers to assessment of student processes and products that directly reflect the learning outcomes (e.g., Stiggins, 1987b). This is in contrast to the indirect assessment of outcomes by asking students questions about their knowledge of the content, as is commonly the case in paper and pencil testing. Performance assessment of such learnings as safety skills in the laboratory would require observation of student behaviours in the appropriate context, as opposed to testing for student knowledge of laboratory safety procedures using a set of written test questions. Student products would include such things as reports from actual library research, and would be directly assessed in terms of completeness of the search, accuracy of the information, and appropriateness of the sources to the purpose of the search. The typical method for assessment of student processes is observation by the teacher (possibly with the aid of rating scales, checklists, etc.), although other approaches are possible (such as self- or peer-evaluation). The method commonly used to assess student products is a type of rating system that is employed by the teacher to assess the quality of the product. An example of this is the marking of a laboratory report where the report is the important outcome (in contrast to assessing such skills as ability to control variables in an experiment, which would probably require a combination of observation and interview).

The teachers had relatively few documents which indicated their performance assessment procedures so detailed analysis was not possible. The review that follows is based on the classroom observations, the teacher interviews, and, to a limited extent, on the documents that the teachers did present (these were typically assignments of projects and papers with some indication how they were to be evaluated). The assessment of processes and behaviours is separated from that for student products.

**Assessment of processes and behaviours.** A number of process skills are identified as objectives in the curriculum, and generally accepted as important by educators. Several examples of process skills are research skills in both social studies and science (e.g., specifying the problem and possible hypotheses; observing, recording and reporting observations; drawing conclusions and making inferences), and laboratory skills in science (e.g., preparing the equipment correctly, such as mounting a slide, focusing the microscope, preparing materials so that there is little contamination). A number of behaviours are also identified: for example, careful and safe handling of all equipment and materials, cleanliness of work station, and cooperativeness and willingness to listen to others. This category can include objectives from all three taxonomic domains, cognitive, affective, and psychomotor, but focuses on the actual behaviour of the students rather than their responses to a set of tasks or questions.

The four teachers did comment that these types of learning objectives are important and that they strive to achieve many of them. However, the appropriate student traits and characteristics typically were not directly assessed. Rather, when they were assessed it was done indirectly--usually by inference from the quality of the product that students were asked for in an assignment. From the classroom observations it was apparent that, for example, proper handling of laboratory equipment was taught by the teachers (how to use microscopes or to mix chemicals), and student practices in the lab were monitored. The quality of the skills and behaviours usually was not rated, and, unless something fairly dangerous or serious occurred, no information based on observations would be recorded. (There were exceptions to this, however, and in one laboratory situation that was observed safety practices were observed and rated by that teacher.). Acquisition of the appropriate skills was inferred if the reported observations were reasonable and the conclusions correct. Some of the exercises were designed so that this inference was reasonable, such as the animal would not be identifiable without proper focusing of the microscope, but in others inference could not be made to various important processes. Similarly, in social studies classes some of the projects and papers were assessed with respect to a number of research skills, these being inferred from the quality of the written report. But there was little or no direct assessment of skills.

The teachers evaluated students' behaviours in an ongoing and general way, and inferred from this characteristics such as level of student motivation, attentiveness, cooperativeness, and so on, but this was usually not done in a systematic and explicit way. One characteristic that was considered important is task completion, and all teachers kept a detailed record of student assignments. This became for the teachers a basis for feedback to the students' parents, as well as providing information to the teacher for determining subsequent student classroom activity and homework. In this sense, then, some of the processes indirectly form part of student grades and reports. Some aspects are reported directly and separately from the achievement grade. This is usually a subjective, global rating (e.g., student is awarded a "4" on a 1-5 scale for "effort and endeavor"), or may simply be an anecdotal remark (e.g., ". . . is inattentive in class", or ". . . participates well in class").

**Assessment of student products.** Student assignments usually involve some product, typically a written paper. All four teachers reported that they required longer assignments of their students, and in some cases, these kinds of tasks can assume a large portion of student grades in some units and topics (e.g., research projects). The teachers permitted, and often encouraged, students to use other forms of presentation in their reports: diagrams, charts, graphs, pictures. They also encouraged use of actual physical models or replicas of phenomena in certain assignments (e.g., scale model of a volcano). Alternate modes of reporting were also encouraged, including visual and oral presentations, and students have used audio- and video-taped presentations in some classes. Oral presentations of reports to the class is a common practice. The teachers indicated that evaluation of the oral reports was usually done by the teacher, and typically involved rating on the basis of content and organization, but may also include matters of presentation such as quality of speech, pacing, awareness of audience, and language use.

Students were sometimes involved in the process of evaluating their own and each others' products. Several teachers noted that they had students rate student presentations, but this was infrequent, and no such instances were observed.

The teachers often assessed the quality of language usage in students' written products. Aspects such as sentence and paragraph structure, grammar, and spelling were considered important by the teachers, and were frequently evaluated as part of the overall mark given to the paper. The simplest way in which this was accomplished was through procedures such as "marks will be taken off for spelling errors". Students were informed as to what characteristics are being considered usually as part of the directions for the paper (e.g., "Answer in complete sentences").

The outcomes assessed in student products are primarily related to the subject discipline: for instance, answers to direct questions are assessed as to their correctness and completeness. In more open-ended assignments, the choice of content and its accuracy were also assessed, but process aspects such as organization, clarity, and neatness, may be assessed as well. The teachers presented the criteria on which they based the assessment to the students for larger assignments, and often these were repeated orally. For smaller assignments the criteria may be given orally and the students were expected to know some of them from previous similar assignments. Some teachers expressly taught students how to answer questions like those that appear on tests and assignments--what constitutes a "good" answer to a given question (e.g., what would be a reasonable definition of some term given the subject matter in which it was to be used). Several of the teachers have produced extensive guidelines for certain kinds of papers and projects. These included substantive concerns and general marking criteria, as well as non-content criteria (such as format, organization, language, etc.). In most cases a format was provided by the teachers for an assignment with indication as to how strictly it was to be followed: fairly strictly in the case of laboratory reports, but simply an example for certain kinds of tasks such producing a title page for a unit of study.

The teachers mentioned the importance of some creativity on the part of students in the presentation of their work. This is an elusive criterion and often was not clearly expressed, but students appeared to be rewarded for doing something somewhat different in a project, and for using ingenuity in the format of the report and making it attractive. This was a major factor in some assignments, such as the creation of notebook covers or title pages. The actual assessment is subjective and typically involved a holistic rating on the part of the teacher (e.g., "three marks out of five for attractiveness").

Assessments of student products were typically included as part of grading and reporting. They could account for as much as 30-40% of student grades for a particular unit or reporting period. Usually one overall mark was given for the product which combines marks based on all of the criteria specified . Thus, insofar as process skills are reflected in the criteria, some would be included as part of student grading.

**Concluding comments**. The teachers made substantial use of performance assessment. The bulk of this assessment was evaluation of student products, and very little involved direct and systematic evaluation of student behaviour. Much of the assessment was based on subject matter, content, although some of it involved characteristics of presentation (e.g., organization, format, style, neatness, attractiveness), aspects of language usage (e.g., structure, grammar, spelling), and elements of creativity (e.g., ingenuity and unusualness in format). Inference was made to student process skills, usually from the product, but sometimes from direct observations of student actions. This was used in planning classroom activities both for individual students and for the class (e.g., particular laboratory skills may be retaught if problems occurred in laboratory reports). Subjective rating schemes were commonly used to mark student papers, and the rating criteria were usually presented to the students.

The teachers occasionally used rating schemes and checklists in their observations of students, but observation for student evaluation purposes was infrequent and informal. Anecdotal forms of recording unusual events and behaviours were the norm, and there was typically no plan for observing specific students in given settings, or to observe all students at one time or another on a systematic basis. Teachers require experience in using systematic methods for the observation and recording of student behaviours, and this should include for purposes of assessing process skills as well as matters of deportment. Materials are readily available to assist teachers in the development of procedures for this (e.g., Cartwright & Cartwright, 1985; Nitko, 1983). These methods should include procedures for systematic identification of students for observation so that information is collected on all students, yet the procedures must be flexible and easy to use and feasible given the realities of large classrooms and limited teacher time (e.g., time-sampling is one possible technique). Teachers should also have the knowledge and skills to specify those student traits most necessary to assess by observation (e.g., which science processes and social participation objectives), and to set forth behavioural criteria and methods to assess them (e.g., rating scales with meaningfully identified anchor points). Teachers also need effective procedures to involve students in the assessments, both to specify the skills and criteria and to employ the assessments on themselves and other students, thereby giving students the opportunity to understand what goals and objectives are expected and how learnings must be displayed. Suggestions to teachers are not as readily available for this.

Teacher skills in the assessment of student products also need to be developed. Teachers should have clear rational for the inclusion of characteristics in the marking scheme, particularly those which are ancillary learning outcomes of the course and program (e.g., language mechanics, neatness). These can be incorporated, but there are situations when ancillary objectives may hinder acquisition of other learnings and become the focus of too much student time and attention. For example, premature or too great an emphasis on accurate and neat recording of observations may conflict with hypothesizing and designing an experiment, or proper paragraphing and sentence structure may detract from conducting a meaningful library search or writing an explanation of a phenomenon.

## Oral Questioning by Teachers

Teachers frequently use oral questions during their classroom instruction (e.g., Stiggins, 1986a; Haertel, 1986). Students may or may not be expected to respond orally to these questions. Oral questioning may be used to elicit a response that is to be assessed but could also be used to motivate students, to gain student attention, to focus students on an important notion, or to help students pose questions that will guide them in knowing whether they understand the notion involved.

Oral questions can be used to find out if students know an answer, what they understand about a topic, or how they feel and what their opinions are. The format can be informal, as with casual questions, or formal, as in structured interviews and oral tests. The questions can be structured or can be extemporaneous based on a general theme. Two types of oral questions are discussed below: those occurring during instruction and those which occur in a testing situation.

**Oral questions asked by the teacher during instruction**. The four teachers frequently presented oral questions during instruction, although they varied considerably in the amount of class time devoted to this format for instruction (it is difficult to generalize but the science teachers appeared to spend much less time on direct instruction than did the social studies teachers, and more time on student assignments). Not all of these questions were for assessment purposes. However, the responses of students were considered important in most cases, and typically were evaluated by the teacher. Sometimes this was by a direct statement of "correct" or "incorrect", but more often it would be followed by more substantial and informative feedback. The teachers stated that they preferred to give positive rather than negative feedback. One teacher in particular rarely told a student that an answer was incorrect; rather, he would use the response to initiate elaboration either by himself or by asking other students. The teachers typically waited for student responses and allowed time for most students to formulate a response before asking for the answer. Most answers were given orally but occasionally students would be asked to write the answer.

The type of questions that teachers ask vary vastly. Some questions require that students simply repeat a piece of information that had been presented. Others ask students to supply examples, to give reasons for a phenomenon, to extend upon an incomplete analysis, or to identify the principle that underlies a phenomenon. One teacher used oral questioning procedures to teach students how to answer questions, by identifying what was a good answer to a question and how a response could be improved.

The teachers indicated that they tried to involve all students as possible respondents to oral questions, and they frequently identified the student that was to respond. The teachers used various techniques in their selection of respondents, some of which were based on a systematic attempt to involve all students (e.g., selecting respondents by starting at one corner of the room and systematically moving across the room). Other bases included selecting students of differing abilities to answer more and less difficult questions, thereby increasing the opportunity of students answering correctly and receiving positive reinforcement. From the classroom observations it was clear that all students were potential respondents. The teachers did not keep written records of either their questioning pattern and strategy or the quality of student responses. Only occasionally would some notation be made.

**Oral testing.** The teachers indicated that they did very little oral testing. They may interview a student, but this would be primarily for purposes other than assessment and would follow some unusual event or behaviour in class, often of a negative nature (such as to determine if there was a problem which prevented a student from completing a homework assignment). The teachers indicated that in some situations and for some students oral testing would be preferred, but one of the reasons given for the paucity of oral testing was the amount of time it would require. One teacher noted that on occasion he would test orally, the reason being that the student did not have the proficiency to read or respond in writing.

**Summary.** Oral questioning during instruction was common, but the purposes it served were varied, and primarily not for assessment purposes beyond that of providing immediate feedback. Teachers attempted to include all students as respondents and were sensitive to potential problems. However, written records of questioning were not maintained. One teacher suggested that teachers with little classroom experience would do well to use seating plans, determine the questioning systematically from the plan, and keep a record of who responded and the quality of the response. It could be argued that experienced teachers would also benefit from such a procedure.

The cognitive level of oral questioning varied considerably, from questions requiring knowledge level thinking to those at the higher levels. This appears to support Stiggins (1986a) who reported that elementary teachers use questions at all levels of cognitive functioning during instruction. Oral testing was rarely used, although acknowledged as being valuable in some cases. It appears that here again some systematic approach to specifying what is most important to assess in this way, and a clear plan of how the assessment is to be done and which students (if not all) are to be assessed. For some students this would provide an important opportunity to express themselves more effectively than they could in writing, and also for the teacher to probe deeper in certain areas. It could be a good method for assessing higher-level thinking skills, and possibly for assessing affective objectives. The results of Stiggins, Griswold, and Wikelund (1989), though, indicate that teachers' oral questioning does not typically emphasis higher level thinking. As much as 40 to 60% of teachers' questions were judged to be at the Recall level. The assessments need not become part of student marks and grades, but could be used formatively. This could be something that teachers decide upon when they determine the objectives that ought to be assessed orally.

## Assessment Feedback to Students

Some information was obtained on the nature of oral and written feedback given to students on their work. As well, the reporting systems used in the schools for informing parents, which also serve to inform students, were obtained (the report cards).

### Oral Feedback to Students

The four teachers typically reported to the class how well they thought the students performed, tending to be positive. They elaborated on questions to which students had responded well. They also reviewed the test with the class, usually asking for participation from students, and identified possible weaknesses which they addressed in greater detail. This would be public feedback in the sense that it was given to all members of the class. Occasionally, individual students would be singled out publicly for praise, or for a particularly interesting answer. Individual students were also given private feedback by comments made to a student at her/his desk or by more formal conferences set up with the student after class or at some other time. These more formal conferences were usually only invoked when there was perceived to be a problem.

The teachers gave group feedback and returned the test or assignment upon completion of marking. Usually this was a day or two after the test was administered or the assignment was due. Thus students received some form of feedback shortly after the assessment. The teachers varied in the amount and type of group reporting they did. For example, one teacher displayed on the blackboard the class average and the averages of other classes on the same test. Other teachers gave an oral report of how well they thought the students performed, which may have included the class average and some expectation the teacher had of what the average should be.

### Written Feedback to Students

The teachers frequently gave written feedback in addition to the summary mark to students on their assignments and tests. This type of feedback was less common for quizzes and every-day assignments than it was for major papers or projects and chapter and unit tests. The feedback ranged from holistic statements, often in the form of praise or exhortation (e.g., "good work!" and "this is well done" or "you need to do more work on this"), to detailed responses to what the student had written. The teachers usually attempted to inform students of where they may have erred in an answer or where they might improve. This was often tempered with some positive statement about the work. None of the teachers posted student marks.

In general, it was not possible to determine how systematically detailed feedback was given to individual students, which students typically received feedback, and how the feedback was used by students. This would have required interviews with students, as well as reviews of samples of feedback.

### Reporting Student Progress

Both schools had fixed reporting periods, at which time a report card on each student would be issued to the parents via the student. The time of reporting and the details and format of what a report card contained vary tremendously across schools generally, but junior high schools typically report achievement in each subject the student is taking. The two schools' report cards contained letter grades for each subject and space for teacher comments, but there the similarities ended. One school issued reports three times in a school year, the other four times.

The amount of detail reported was much greater for one of the schools. This school required that a progress report be maintained for each student in each course; this record included the teacher's rating on a 4-point scale of seven aspects (homework, daily assignments, tests, projects, effort, participation, and behaviour) as well as anecdotal comments. The progress report formed the basis for the report card, which included alongside the grade awarded a rating in each course of student effort. Also, this school required a breakdown of marks achieved in the course by the student for the reporting period (marks on chapter and unit tests, assignments, class participation, etc.) and further anecdotal comments on student progress. If the school deemed it desirable, as determined by the principal and all teachers who teach the student, this second record was sent to the parents along with the report card. The teachers thought that parents and students obtained much useful information from this kind of reporting; they were clearly informed of the student's progress. But it should be noted that the teachers found it an incredible amount of work to make meaningful individual comments for many students. There was teacher dissatisfaction with the reporting procedures in the other school and, at the time of the study, they were under review by a committee in the school.

It would be reasonable to generalize that student achievement in the subject forms the basis of reporting to parents. Marks are typically reported as letter grades, but these can be translated to a 100-point scale (usually from which the grades were derived originally). In the two schools, the report cards contained the grade-scale conversion ($A^+ = 90\text{-}100$, $A = 85\text{-}89$, etc.). The teachers indicated that they accumulated the marks for each student across all assignments and tests, and divided by the total possible number of marks. This was then converted to the appropriate letter. The weighting of an assignment was typically a result of the total number of marks given it, although in some cases there may be a factor applied to the marks. Clearly, there are problems with the way in which information was accumulated and grades were computed. This is well recognized in the literature (e.g., Friedman & Manley, 1991; Manke & Loyd, 1991; Stiggins, Frisbie, & Griswold, 1989; Terwilliger, 1989). Differences in difficulty levels and in score variances of different components were not considered, and weighting may or may not be applied appropriately (e.g., Oosterhof, 1987; Thayer, 1991).

## The Practices of Classroom Assessment and What this Means for Teacher Preparation

The teachers in the case studies were involved in a variety of assessment activities, and these were for a number of purposes. In general, the teachers used some form of paper and pencil assessment as the most common vehicle for formal assessment. There may have been more than one purpose for an assessment, but grading and reporting was the primary one. Evaluating instruction was rated as next in importance, and although there was little clear evidence of how this was used, the teachers commented that they did review their instruction and the course more generally as a result of assessments. There was very little diagnostic assessment conducted, although this purpose was considered fairly important by several of the teachers. In part this was stated to be due to the nature of science and social studies, very little exists in the way of diagnostic theory and procedures, but also because at the junior high school level classroom teachers do not usually develop individual programs for students. Assessment information was used to inform instruction, but more for group reteaching and review purposes than for individual student programming.

The quality of the assessment materials was found to generally acceptable technically, but assessments clearly focused on the lower levels of cognitive functioning. This was true particularly for tests and quizzes, which were rated as the most important vehicle of assessment. Also, other aspects of learning in these subject areas, such as science processes and research skills in social studies, were not well represented in any of the assessments. Affective characteristics were rarely the focus of assessments, although an example of where this did occur was the direct assessment of safety practices in laboratory work. Certain affective characteristics were included in assessments but in an implicit manner, such as by penalties being imposed if work was not completed on time. This was made more prominent since in-class and homework assignments, and longer-term projects, were included as part of composite marks for reporting progress.

The preferred formats on paper and pencil tests were supply-type items which required short written responses. These were rated to be, almost exclusively, at the knowledge and comprehension levels of Bloom's taxonomy. These item types were followed in frequency by selection-type items including true-false, matching, and multiple-choice, which were also rated to be at the lower levels of Bloom's taxonomy. These findings are in keeping with those of other researchers, such as Fleming and Chambers (1983), Stiggins (1987a), and Stiggins, Griswold, and Wikelund (1989). Direct assessments of student products occurred in the form of evaluation of laboratory reports, research papers, and projects, but these featured less prominently in students' composite marks. Direct assessments of student behaviours were reported to be used rarely, and were observed on only one occasion. In both of these forms of performance assessment it was difficult to determine whether systematic assessment procedures were followed, although the teachers indicated that they use rating schemes of some type for the assessment of student products.

Experienced teachers do conduct appropriate assessments using a variety of formats, although perhaps not as frequently and certainly not as systematically as some measurement specialists advocate (e.g., Stiggins, 1988a; Wiggins, 1989b). It appears fair to conclude that it is possible and reasonable to expect teachers to develop and use both paper and pencil and performance assessments, and that these should encompass the full range of learning outcomes of coursework in our schools. The preparation that teachers need to do this includes the technical skills associated with developing systematic and appropriate measurement tools, such as preparing good test questions and rating procedures, and the broader understandings of how and when these are to be applied.

It also appears that various technical procedures of the measurement process, such as numerical item analysis and reliability estimation, may be possible to employ in the classroom but only in limited ways. Teachers are more concerned with conducting assessments which appear to meet the objectives they have set for the course, that cover the content they have taught, that are fair and reasonable for their students, that are reasonably efficient in the classroom, and that can be used directly in forming grades for reporting purposes. Although teachers give considerable thought to assessment, this does not appear to reflect concerns from an assessment point of view: for example, reliability and validity do not form guiding concepts for enhancing an assessment. Rather, the content of the course, the beliefs about what students can do, and the perceived importance of various aspects of learning (e.g., certain process skills, language skills, motivation and effort) appear to form the guiding concepts in deciding what to assess and how to assess it.

Any approach to preparing teachers for their classroom assessment tasks must provide them with the technical skills to produce high quality paper and pencil assessments; these form the bulk of their formal assessment. As well, it must provide ample experience in developing and using other forms of assessment, particularly those that assess directly student products and processes. The approach must give teachers the background to determine what is appropriate to assess, and under what conditions and in what settings various assessments ought to be conducted. Teachers must have the ability and the confidence to assess all of the important skills and outcomes identified in our school curricula. Further, teachers must be able to distinguish differences among various purposes for assessment, and know how this affects their evaluation. For, example teachers must know how to separate assessment for purposes of ongoing instruction, both group and individual diagnosis, from summative assessments designed to give overall indicators of student learning.

Finally, the grading and reporting function of assessment cannot be downplayed, and all aspects of this must be thoroughly developed in teacher training. Part of this includes training in forming grading schemes, but perhaps fully as important is development in deciding what is to be communicated by grades or by other means, and how communication can be done effectively and fairly.

# IV. COMPONENTS OF A PROGRAM FOR TEACHER DEVELOPMENT IN CLASSROOM ASSESSMENT

The instructional component for professional preparation and development of teachers for assessment must be both practical in the classroom and consonant with good assessment principles. Four major tasks were required to identify recommendations:

1. Specify the characteristics of good classroom assessment practices.
2. Evaluate present teacher assessment practices with respect to these characteristics.
3. Prepare recommendations for teacher preparation in classroom assessment based on the characteristics of good assessment, and on the realities of the classroom.
4. Review the recommendations for their adherence to modern assessment principles and their practicality to the classroom setting.

First, classroom assessment practices are discussed relative to their purposes and to the context in which they occur. In the second section the characteristics of good classroom assessment are presented, with consideration given to modern assessment theory and present classroom practices. Four categories of characteristics are identified: reliability, validity, utility, and efficiency. Recommendations for the appropriate preparation of teachers for their classroom assessment tasks are presented and discussed for each characteristic. These recommendations are summarized in the third section. The procedures for critically reviewing the recommendations, and the results of the review, are given in Chapter V (with supporting material in Appendices D and E).

## Assessment in the Classroom Context

The characteristics that are considered important to classroom assessment are dependent on the purpose of the assessment and on the context in which the assessment is conducted. Classroom assessment is part of the larger process of teacher decision-making, although only a fraction of teachers' decisions are based on what we typically include under the banner of assessment. Clark and Peterson (1986) concluded that on average interactive decisions are made every two minutes in a classroom. Certainly, one could not expect that the kinds of rapid decisions made by teachers during active instruction would be based on judgements that involve much more than shrewd and experienced guesses by the teacher as to what is meant by students' responses and behaviours during this interaction. However, there are many classroom decisions that fulfill purposes other than those involved in interactive classroom processes and these can be much more deliberate in nature, such as decisions regarding the topics that should receive review and additional instruction at the end of a unit of instruction. It is to these more deliberate decisions that the characteristics of good assessment practices can apply.

### Purposes of Classroom Assessment

There are a number of purposes of classroom assessments and teachers vary in the relative emphasis placed on them, but the literature and the case studies highlighted three as being particularly significant to teachers. The three most important purposes, in order of their perceived importance to teachers, were:

1. Assigning grades to students,
2. Evaluating instruction, and
3. Diagnosing individual and group instructional needs.

Two teachers in the case studies placed less emphasis on grading and reporting and more on diagnosing student difficulties and determining instructional effectiveness. Gullickson (1982) and R. J. Wilson (1990) reported that teachers at the higher grade levels tend to emphasize the grading and reporting function more than those at the lower grade levels, although this difference is slight, and teachers at the lower levels more frequently noting the diagnosis purpose. Dorr-Bremme (1983) also noted variations among teachers, but also reported that the differences in assessment purposes between elementary and high school teachers were slight.

Grading and reporting student progress is clearly one of the most important purposes of classroom assessment at all grade levels, and the most important purpose of formal types of assessment at the higher grades, such as paper and pencil tests. Teachers and schools take seriously the need to judge or appraise student learning and to report this formally. Promotion decisions are based primarily on these evaluations, except in the very early grades where factors such as social maturity may also be considered. Crooks (1988) and R. J. Wilson (1990) noted that teachers at the higher grade levels also frequently report the importance of counting assessments for grades in motivating students and keeping them on task. The case studies indicated that for some teachers virtually every classroom assignment is treated as an assessment and is included in some way in the computation of term and final grades.

The second purpose was endorsed by teachers in other studies as well (e.g., Gullickson, 1982; Webster, 1987), but it is difficult to determine if teachers clearly distinguish this from the purpose of diagnosing group instructional needs. This purpose reflects broader, summative evaluation of the effectiveness of the teacher's overall instructional approach and of the curriculum and the materials used. Teachers in the case studies were questioned in the interviews and appeared clear on this distinction, and three did identify this purpose as important.

In the early grades diagnosis of individual student needs is emphasized by teachers, although this diagnosis is generally in the areas of language arts (especially reading) and mathematics. The junior high teachers in the case studies indicated that much of this kind of work is left to specialized educators, such as resource teachers, and that there was little or no individual diagnosis carried out in science or social studies. These teachers further noted that subject areas other than reading and mathematics are not structured in a way to permit meaningful individual diagnosis. In the literature, teachers reported that assessment was important to planning subsequent teaching for the class as a whole, or for groups of students in the class (Herman & Dorr-Bremme, 1983; Webster, 1987). A simple example of this would be the reteaching or review of a topic for which the items on a test were poorly answered.

Teachers tend to use assessment results for several purposes, so that a test which is counted towards grades is also often used to provide a check on instruction and identify areas of student problems (R. J. Wilson, 1990). The results of most classroom assessments also have multiple audiences. For example, in classrooms beyond the primary level, most assessments, irrespective of their primary purpose, involve feedback to the teacher and to the student, some assessments would be included in formal reporting of progress to parents, and perhaps a few would be used to provide information to the school or to groups external to the school. Many assessments, too, would involve the purpose of diagnosing instructional needs, even though they might be designed more for grading and reporting progress. It could be argued, as does Crooks (1988), that most assessments (particularly those that involve formal and explicit procedures) serve the purposes of communicating achievement expectations and providing students with test-taking experience, although these typically would not be the express purposes of an assessment, as indicated by teachers in the case studies. In a general way, too, most assessments provide teachers with information as to the effectiveness of their teaching and of the program, although there is no strong evidence that teachers adjust their teaching to any great extent on the basis of their assessments.

R. J. Wilson (1990) noted an additional purpose for which assessment was used: that of providing students with practice in applying what they were learning. This purpose was not highly endorsed by teachers in the case studies.

The purposes that are primarily considered in identifying the characteristics of classroom assessment are assigning grades and reporting progress, diagnosing instructional needs, and determining program effectiveness.

## The Context of Classroom Assessment

Deliberately structured assessments provide teachers with information that assists them in making long-range decisions regarding instructional effectiveness and student progress, but may also provide information that assists them in forming more accurate rapid-fire decisions during instruction. To do this the assessments must reflect, and be sensitive to, the ongoing instruction in the classroom and the conditions under which the learning takes place. This has been referred to as the fidelity of the assessment, and is discussed under the topic of validity in the next section. Context refers to the environmental factors affecting both instruction and assessment; that is, the social and community setting of the school, the structure and climate of the school, and the composition and dynamics of the classroom. The context is the social milieu of the classroom that provides both support for instruction and restrictions on it. For example, the nature of the community gives rise to certain expectations for both the learning and the behaviour of students. One community may be very demanding of its school and may support high standards of performance, whereas another may have very little involvement with its school. This directly affects what teachers can do in the classroom and the demands they can put on students. There are some schools where the teachers cannot expect any assigned homework to be done, there is not the support for it in many cases, and students simply do not complete the assignments.

It is not possible to identify all, or even a large portion of, the factors that impinge on classroom instruction. However, there are some variables that are clearly related to classroom assessment and to its validity for various purposes. These can be broadly grouped as in-class, school, and extra-school factors. In-class factors include grade level, course or subject area, size of the class, heterogeneity of the class, class dynamics, and ability levels of the students. School factors are such things as materials, facilities, and equipment available; teacher support available (e.g., resources, teacher assistants, preparation time); climate of the school (e.g., administrative leadership); and policies and general focus of the school. The extra-school factors could include community support of the school (e.g., financial, moral), ethnic make-up (e.g., language differences), expectations of the school (e.g., standards expected, modes of behaviour demanded), and political demands on the school.

Although not all the of factors listed above impinge directly on classroom assessment practices, many of them have an indirect effect on what and how teachers assess their students. For example, a community may support high academic standards in a school, but may not tolerate the inclusion of affective assessment in the grading and reporting process. The community, too, can have considerable influence on how student progress is reported. Some communities simply will not accept narrative reporting procedures. Messick (1989a, 1989b) notes, indeed emphasizes, that the social-consequential implications of assessment should not be ignored. Assessments of students have meanings and they certainly have effects on students, on parents, on teachers, and on the system. These effects may well lead to pressures on what happens in classrooms and on assessment practices. However, the school context is complex and dynamic, and its effects cannot be readily identified and predicted.

## Characteristics of Good Classroom Assessment

Good classroom assessment practices must be considered in relation to the complex milieu of teacher and student interactions. What is possible in a classroom is a far cry from what might be possible in a less hectic environment and from what would be deemed ideal by measurement specialists. Furthermore, a teacher's world is a busy one, and the tasks of preparing, conducting, and marking students' work, and recording and providing feedback must be balanced against the tasks of planning and preparing for instruction, organizing and managing classrooms, meeting and working with students, and so on. Therefore, the characteristics of good classroom assessment practices must reflect the realities of the classroom setting. The realities of the classroom are reflected in the report on the case studies (Chapter III). The case studies were based on experienced, recognized teachers, so should be indicative of what is reasonably possible.

In this section the characteristics of good classroom assessment practices are identified from the perspective of measurement and evaluation theory. There are many technical aspects of measurement and evaluation that can be applied to classroom assessment practices, although Cole (1987) argues that these technical aspects are more the characteristics of "assessment designed for measurement" (external assessment) than of "assessment designed for instruction" (classroom assessment). However, the general characteristics of measurement are applicable. They can be identified as reliability, validity, utility, and efficiency. The first characteristic encompasses the traditional concepts of reliability and error of measurement. The second, validity, is based on Messick's (1989a, 1989b) unified concept of validity, and includes content/curricular, criterion-related, and construct validation, as well as notions related to the social consequences of testing, notions of authenticity, equity, and fairness. The third, utility, is strongly related to the purpose of the assessment, and includes the notions of referential basis, discrimination, communicability, and effectiveness of the scores. The last characteristic, efficiency, refers to the practicality and ease of use of the assessment for classroom purposes.

Consistent with Angoff (1988) and Messick (1989b), it is necessary to state at the outset that the preeminent characteristic of any assessment process is validity. The other characteristics are subordinate to validation of the assessment. For example, if the results of an assessment are to be valid, they must be reliable in some way. To some extent, the purpose of the assessment prescribes the validation evidence required for its appropriate use. Further, the characteristic utility is often considered to be an issue of validity (e.g., Messick, 1988). However, to more clearly identify the interpretive aspects of assessment use this characteristic was separated. Other breakdowns are clearly possible, but these were chosen as convenient categories to reflect the kinds of concerns highlighted by each, and the implications for classroom assessment that spring from these concerns.

Criteria can be identified for each of the characteristics identified. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1985) gives general guidelines and criteria for the production of tests, although many of the guidelines apply more to commercial and standardized tests than to classroom assessments (Frisbie & Friedman, 1987). A number of measurement textbooks provide guidelines and rules for the construction of teacher-made tests (e.g., Ebel & Frisbie, 1991; Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Nitko, 1983). Textbooks generally reflect the measurement lore of the authors, and incorporate psychological or measurement theory for various topics, particularly those involving psychometrics. Commonly used educational measurement textbooks have been criticized by Gullickson and Ellwein (1985) and Stiggins and Bridgeford (1985) as having too great an emphasis on psychometric theory, statistics, and standardized testing, and having insufficient coverage of assessment procedures more appropriate to the needs of teachers in classrooms. However, some guidelines provided in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1985), in textbooks such as those listed above, and in various documents related to classroom testing (e.g., Stiggins, 1987a), were deemed applicable and were used in setting criteria for teacher assessment practices and in developing the recommendations for teacher development.

## Reliability and Error of Measurement

Reliability attempts to answer the question, Are the evaluations based on accurate, replicable information? The term reliability is used here as it is defined by Sirotnik (1983), which is similar to its use in classical true-score test theory (such as is described by Thorndike, 1988). It refers to consistency or agreement over repeated measurements, be they across occasions, forms of measurement, observers, or some other method of obtaining multiple measures of the same construct. Reliability is to be distinguished from the notion of accuracy as correctness of decision-making. This is the familiar distinction made between decision-consistency and decision-accuracy in criterion-referenced measurement by authors such as Crocker and Algina (1986). Reliability is used in its consistency sense since decision-accuracy with the attendant meaning of correctness is more a matter of external validity.

Standard error of measurement is another way of identifying and quantifying measurement error, which is inversely related to reliability. Standard error summarizes the within-person inconsistency in score-scale units of the measurements (Feldt & Brennan, 1989). This highlights directly the error about a measurement, and thereby emphasizes the effect of unreliability. However, it is a statistic that is not free of the scale of the scores, and therefore is more difficult to apply to various sets of measurements and use to judge the quality of assessment procedures.

Reliability, in true-score theory, encompasses a number of procedures for estimating the stability of a measurement, each procedure taking into account one or more of several sources of error. The errors are usually assumed to be independent of true scores and to be randomly distributed. It is the accuracy of the measurements that is determined by reliability, and it is not a characteristic of the measuring device: reliability is a property of the scores and not the test instruments (this is commonly stated, even in measurement practitioner-oriented texts such as Gronlund, 1985). Reliability is used here to refer specifically to measurement in practice, and not to the scale itself nor to the theory underlying the derivation of the measuring scale. The theoretical basis of the scale is more a matter of construct validity (internal validity). Reliability is restricted to the more precise notion, whereas some authors (e.g., Anastasi, 1988; Ebel & Frisbie, 1986) include internal consistency as a type of reliability, which Sirotnik (1983) argues is a quality of the scale itself. Reliability, then, is equivalent to accuracy, which distinguishes the notion intended here from that of precision, which also is a characteristic of the measuring device. Accuracy, as the term is used in the physical sciences, implies the need to consider the scale and range of scores in the interpretation of score error. It is an empirical concept reflecting the amount of error involved in taking measurements relative to the size or range of the measurements taken. Feldt and Brennan (1989) note that the best estimate of the reliability coefficient is one that includes all sources of measurement error, although this is clearly not practicable.

An alternative to the term reliability is to use generalizability, which is defined by Cronbach and his associates as accuracy of generalization (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Generalizability encompasses the traditional notions of reliability, and readily subsumes the various types of reliability as well as the methods for obtaining reliability estimates under the more general framework of identifying the expected error for a given plan (universe) of generalization. It also allows, in fact implies, that there is no "true score" as such but as many universe scores as there are universes of generalization with differing sources of error and error interactions. Other than the assumptions of true scores and strictly parallel test forms, generalizability theory makes much the same assumptions as do the earlier theories of reliability (e.g., independence of errors from true and facet scores and from each other, such that a linear model can be applied), and thus it could include internal consistency as one form of reliability. Also, as noted by Thorndike (1988), it can include some notions that we commonly attribute to the concept of validity. The facets of a study to determine generalizability could readily incorporate factors across which differences could exist theoretically, and generalization in this sense is really a substantive part of what Cook and Campbell (1979) call external validity. Generalizability theory has been extended and adapted to criterion-referenced testing. The dependability of domain scores (analogous to generalizability of norm-referenced scores) suggests an index that uses the criterion or mastery cutoff as the reference point for calculating variance estimates rather than the universe or population mean (Brennan, 1984; Brennan & Kane, 1983). However, there are no reports indicating whether or not this approach has been used successfully in criterion-referenced test applications.

Since the term reliability is the one most commonly found in the applied measurement literature, and the one teachers are most likely to encounter and be expected to interpret, it was used. It has stood the test of time, despite numerous developments and possible imprecision in meaning.

The notions of consistency and reliability underlying both criterion- and norm-referenced applications are applicable to classroom assessment. Studies by R. J. Wilson (1990) and Stiggins, Frisbie, and Griswold (1989) indicated that teachers do not use the systematic procedures recommended by measurement experts for either criterion- or norm-referenced approaches to evaluation. Instead, the teachers used methods to obtain the composite index of achievement that included subjective guesses and imprecise weightings of various tests and assignments ( e.g., Manke & Loyd, 1990, 1991; Stiggins, Frisbie, and Griswold, 1989). Teachers in the case studies indicated that they often attempt to accommodate particular students by including on a test some easier questions ("so everyone has a chance to get some answers correct") and some more difficult questions ("to challenge the better students"). At the lower grades change or growth on particular skills is used by some schools to determine student progress. Therefore, teachers should be familiar also with the theory and problems surrounding self-referencing and the reliability of difference scores since they may encounter school systems that base student achievement and progress reports on change and growth.

Much has been made of the distinction between norm and criterion referencing and how this distinction relates to reliability (e.g., Berk, 1984a). Teachers appear to be attracted to the notion of criterion referencing, and even of self referencing, but what little research there is suggests that they use a common-sense combination of norm and criterion referencing often with some adjustments made in the assessments to accommodate students in the class who may be particularly weak or strong. Although generalizability is the preferred approach to reliability estimation in a norm-referenced context (e.g., Shavelson & Webb, 1991) Feldt and Brennan (1989) suggest classical reliability theory based on parallel forms provides sufficient guidance for the identification of norm-referenced reliability procedures that might be appropriate to classroom assessment. Berk (1984a) concluded that for criterion-referenced applications Hambleton and Novick's (1973) $p_0$ index "appears to have the greatest utility for classroom test construction and decision making" (p. 260) and further suggests that it be calculated from two administrations of the test. This index indicates the proportion of decisions (e.g., mastery-nonmastery) which are in agreement from the first administration of a test to the second. Martuza (1977) and Nitko (1983) suggest using Cohen's kappa statistic ($\kappa$), which is a correction of $p_0$ for chance agreements, although Gronlund (1985) recommends using $p_0$ for classroom criterion-referenced applications since it is more readily computed and understood. Both the norm- and criterion-referenced approaches to reliability are appropriate to test-retest applications as well as to those involving alternate forms of the test. They can also be applied to situations involving ratings by multiple scorers or observers, although this blurs the distinction between reliability of measurements and observer agreement: observer agreement is a necessary but insufficient condition for reliability (Frick & Semmel, 1978).

In spite of the appropriateness of the above reliability estimation procedures to classroom assessments, and of the simplicity of the $p_0$ and $\kappa$ indices, they may be impossible to use in everyday classroom settings. It is extremely unlikely that teachers will administer a classroom test more than once to the same group of students without intervening instruction, or that teachers will produce and administer parallel forms of a test to the same students. Also, it is rare for more than one teacher to observe and assess the same student behaviours or products, although there are occasional and important instances where ratings by two or more teachers are obtained, such as in large-scale holistic scoring of writing. Testing consumes considerable amounts of instructional time, as well as teacher out-of-class time, so it is difficult to justify two administrations of the same test or multiple ratings in a classroom setting. To expect teachers to do this is unrealistic, but particularly so for tests that are of great consequence to the students. These *high stakes* tests (as this is described by Madaus, 1989, and earlier by Popham, 1980) usually form the basis for grading and reporting. They are given only once to a particular group of students, so it appears necessary, then, to obtain a measure of accuracy based on only one administration of the assessment device or procedure.

Reliability for classroom assessments cannot be restricted to one estimation procedure as there are several purposes for which teachers assess their students, and many different assessment methods are used. However, some guidelines can be put forth for some of the assessment procedures used by teachers. These guidelines must reflect two realities. First, it is extremely unlikely that teachers will obtain more than one set of scores using one assessment procedure (teachers just do not have the time, nor is it something that works well in a classroom situation). Second, teachers do not use anything beyond the most simple of statistical calculations (e.g., Gullickson & Ellwein, 1985; McLean, 1985; the case studies), so any numerical procedures must be easily applied, computed, and interpreted. But teachers must be clearly aware of the characteristic of reliability, its relationship to the assessment purpose, and the implications for practice. Teachers must be clearly aware of the factors which affect error in measurement and practical procedures for reducing this error. Also, some of the numerical procedures for both criterion- and norm-referenced scores can be made simple enough so that in situations where reliability is crucial (high-stakes tests such as final examinations) they can be used.

**Recommendations for reliability.** There are a number of recommendations to be made regarding reliability. These are based on what is practicable in the classroom as well as on what we know of present practices. Reliability and error of measurement must be understood by teachers in relation to the purposes for their assessments. The more consequential the evaluation, and the less reversible the decision to be made, the smaller should be the error in the measurements that are used to inform the evaluation. Teachers should know how reliability can be improved for common classroom assessment procedures, and how it relates to various common classroom assessment procedures.

Several criteria can be applied to classroom assessment regarding reliability. These are noted in the discussion following each recommendation.

1. The importance of reliability is directly dependent on the consequences of the assessment. In classroom assessment, reliability should be greatest for those assessment results having the most impact on the students, high-stakes assessments. Therefore, teachers must be able to identify the main purpose of an assessment and the personal and social consequences for the student. This is necessary also for secondary purposes if these exist. To understand the importance of reliability for high stakes assessments, teachers must be aware of the implications that low reliability and errors of measure might have for various decisions.

It was evident from the observations and from the review of teacher-made tests in the case studies that teachers exhibit more care in the development of their formal testing procedures than they do for less formal assessments. At grades beyond the primary level formal tests comprise much of the basis for grading and reporting student achievement, although Webster (1987) reports that this is by no means consistent across teachers, subject areas, and schools. Reliability is important for formal testing, but there is no evidence that teachers develop and administer their tests with a view to enhancing reliability. Teachers do not typically identify in advance, or systematically, that which they deem to be most important to assess (a validity issue) and insure that they have adequate measures of students' performance on these important aspects. As well, most teachers include classroom assignments and other activities as part of assessment for grading. These assignments are not as carefully designed as are most tests, and involve procedures that have even lower reliability.

Teachers' assessments are frequently used for multiple purposes (e.g., both for grading and to provide instructional diagnosis). All purposes for an assessment should be clearly identified, and this should guide teachers in the design of the assessment procedures. If grading is one of the purposes of a test, the test should reflect in a balanced way the topics taught and with sufficient measurement so that the overall score is reasonably reliable. Further, if instructional diagnosis is one of the purposes, then there ought to be sufficient assessment(s) of each topic or skill which is to be diagnosed. There is no reported evidence as to whether or not teachers do this, but there is evidence from the case studies and from Webster (1987) that teachers do not formally set out a test design, such as a table of specifications or test blueprint, that would make explicit what is to be assessed and the amount of assessment that is afforded each area or topic covered by the test.

Some purposes are incompatible in one assessment procedure, such as detailed individual diagnosis and comprehensive summative testing, or assessment that occurs as part of learning activities and assessment at the end of instruction. There is insufficient detail in the summative test to provide reliable diagnostic information. It is also inappropriate to include all assessment information as part of summative grading and reporting, as some of this does not reflect students' achievement since it is based on assessment during the learning process and reflects incomplete learning.

There are no formal guidelines regarding the desired level of reliability for various purposes, if the reliability were expressed as a numerical index. However, some authors do provide suggestions for particular test types. This is discussed in point 4 below.

2. The reliability of assessment information can be enhanced in two general ways which are practical in the classroom: by making the assessment procedures more explicit and systematic, and by increasing the amount of systematic and appropriately collected information. Teachers must know how these apply to particular classroom assessment practices. Teachers must know how to enhance the reliability of various types of assessment that are appropriate to the intended learnings, such as subjective observations, long range assignments and projects, constructed response testing, objective testing, and portfolio assessment. Each of these may have unique problems of measurement error, but usually error can be reduced by applying the two general ways noted above.

According to Fleming and Chambers (1983), Stiggins (1989), and the case studies, directions to students on teacher-made tests are frequently not clear, including the number of marks awarded to various items. This can make it difficult for students to respond appropriately, leading to lowered reliability in the scoring of constructed responses (as well as jeopardizing validity directly). The wording of test instructions and of the items and tasks themselves must be clearly understood by all students. Clarification of instructions, and of item wording, can be accomplished simply by having other teachers review the assessment in advance, providing feedback, and by asking students for feedback (as noted by the teachers in the case studies, students often welcome the opportunity to comment on a test, and they can be very perceptive). This also provides the basis for justifying marks awarded for student responses and for students to judge their responses and to question the marking.

The reliability of scores obtained using one test format may differ from those using another. Certain procedures are prone to more error: for example, two-choice items (such as true-false) usually yield less reliable data for the same number of items than do four-choice items (Frisbie, 1992). However, the reliability of most assessment formats can be enhanced by improvement in the clarity of the items themselves (e.g., clearer wording) and by increasing the number of items or amount of assessment. Teachers should be aware of the threats to reliability for commonly used classroom assessment techniques and item formats.

The procedures for conducting an assessment that involves constructed responses, either written or in some other format, should be systematic and structured, insofar as this is appropriate to the specific course objectives. Some important learning outcomes require that less structured procedures be used (e.g., Wiggins, 1989b; Wolf, Bixby, Glenn, & Gardner, 1991), in which case students should have ample experience with these procedures. Students must have a clear set of objectives in doing the assessment otherwise they may respond in inappropriate ways making the responses difficult to evaluate accurately (and fairly). The marking procedures should be clearly identified in advance to students, and it is useful to prepare a model answer to use as a benchmark against which to judge student responses. These procedures are clearly outlined in introductory measurement textbooks (e.g., Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Popham, 1990). There are numerous procedures available today for marking written work systematically (e.g., Huot, 1990, reviews holistic, analytic, and primary trait scoring; Biggs & Collis, 1982, outline the SOLO taxonomy), but for many of these procedures it is advocated that more than one marker reads the paper. This is not usually practical in the classroom on an everyday basis, but it is possible for teachers to calibrate their marking by procedures such as repeat marking of randomly selected papers and by exchanging high-stakes written assignments with other teachers on a regular basis.

Salmon-Cox (1981), Stiggins and Bridgeford (1985), Webster (1987), and the case studies all indicated that teachers frequently use observation as a means of gathering day-to-day assessment data. Observation is also an important part of applied performance assessment. For observation procedures to be reliable, it is necessary that teachers use systematic procedures for identifying what is to be observed, the conditions under which the observations are to be made, and how the observations are to be judged, much as in designing and marking constructed response assessments. Furthermore, it is impossible to observe all students simultaneously, or to observe all the important behaviours of any one student. A systematic procedure must be invoked so that some behaviours of all students are observed at some time under the appropriate conditions (procedures such as time sampling observation segments and randomly ordering students for observation). Procedures for this are not as well developed as they are for assessing students' written work, but some guidelines and suggestions are given in modern measurement textbooks and sources such as Berk (1986), Cartwright and Cartwright (1984), Fairbairn (1988), and Stiggins (1986b).

The second way of enhancing reliability is by increasing the amount of systematically and carefully gathered information on the student regarding the particular skill (e.g., increasing the number of test items, using two assessments, making more observations). This presupposes that each unit of assessment information has some inherent reliability, and that there are positive correlations among the assessment units.. This is based on the general principle that decisions based on more information are likely to be more stable than those based on less, provided the pieces of information have similar levels of error. The principle also underlies the generalized Spearman-Brown prophecy formula of classical test theory. Increasing the amount of assessment or number of independent measurements is particularly crucial for high-stakes decisions, and in cases where any one assessment is likely to be unreliable (e.g., subjective assessments, observation-based assessments). Students should also have the opportunity to practice and become familiar with various approaches to assessment, particularly if not all the procedures are generally familiar to all students.

**3.** Students must be given the opportunity to exhibit their skills on several occasions, particularly if there is some doubt as to the level of performance of the individual or if the skill is mastered. This does not mean that simply "more is better" in terms of assessment, as may be suggested in the preceding section. More assessment of the same ilk may yield spuriously high reliability estimates since only one potential source of error is included. The true reliability of a composite score is not enhanced by simply increasing the number of scores or measures taken into the composite, but rather by increasing the occasions on which students can exhibit their learning. In part, this reflects the principle that the assessment units or measures must have some reliability and must be intercorrelated, but more the principle noted earlier by Feldt and Brennan (1989): that as many sources of measurement error as is meaningfully possible should be taken into account in estimating the overall reliability or error. Clear potential sources of measurement error in classroom assessment are procedures and formats of the assessment, particular situations of time and setting, and any of a myriad of individual student factors such as health and personal relationships. Teachers must be able to obtain assessments of skills and knowledge using various appropriate procedures and in different settings and occasions.

Discussions with and observations of teachers in the case studies suggested that teachers do provide multiple opportunities for students to exhibit their learning, at least in an informal way. They often ask that students do questions again, for example. Dorr-Bremme and Herman (1986) quoted a teacher in this regard: "You can't count on a score on one test too heavily. The kid could be sick or tired or just not feeling up to doing it that day. Maybe his parents had a fight the night before. Maybe he doesn't test well" (p. 43). Often assessments are repeated more for remedial purposes rather than to confirm previous findings. There is considerable error about any score, and care should be taken that more information is sought in cases where students are borderline, or where scores are close to a cutoff (for grades, or mastery designation). But this additional information must be based on appropriate systematic assessment procedures, and not simply on quick questions or incidental classroom information.

**4.** There are a number of numerical procedures that can be used to help determine if classroom measurements are reliable. Some of these are fairly simple to compute, and, since modern computing equipment is available in most schools, it is becoming reasonable for teachers to check the reliability of their tests periodically. Teachers ought to confirm the adequacy of their measurements and check the reliability of some of the more high-stakes assessments using straightforward numerical procedures.

Teachers rarely obtain multiple measures using the same or very similar instruments; as noted above, this is not usually practical in a classroom setting. Therefore it is necessary to obtain reliability estimates based on single administrations of an assessment. Reliability can be important at the item, subtest, or test level, depending on the use of the assessment. Analysis of student performance on particular skills implies that there should be more than one question (or more than one independent measurement) for each important topic or skill being assessed. If performance is interpreted in a criterion-referenced manner, scores on these parallel measurements can be compared using a statistic such as $p_0$, which is readily computed and interpreted. It may be possible to compute the slightly more complex, and preferable, statistic $\kappa$ in situations where there are teachers with experience in statistics (such as in some of the larger high schools). If the scoring procedures provide scaled or ranked scores, it may be more appropriate to compute a correlation coefficient such as Spearman's $\rho$ or Pearson's $r$. Correlation coefficients can be used in criterion-referenced situations to estimate reliability of particular items and test procedures provided that these procedures provide some distribution of scores. These statistics are described in classroom-oriented measurement textbooks, such as Gronlund and Linn (1990) and Nitko (1983), but are often considered too cumbersome for classroom teachers to employ on a regular basis.

If the purpose of an assessment is to provide diagnostic information, it is necessary to obtain reliabilities for parts of the assessment, such as subtests (the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1985) recommends computation of subtest reliabilities if scores on subtests are to be interpreted). As with the analysis of particular assessment items, this would involve an inordinate amount of work for teachers if they are to compute reliability coefficients for subtests on many of their tests. Also, teachers do not usually make systematic, high-stakes use of subtests in their testing. Except in limited situations, it is unreasonable to expect the calculation of subtest reliabilities, even if these require only simple calculations such as those necessary for $p_0$'s. However, it may be possible for schools and systems to assist teachers to obtain item and subtest reliabilities for selected situations.

For many classroom tests students' scores can be expected to distribute themselves across a substantial portion of the possible score range, and norm-referenced procedures could be applied to estimate reliability of any one assessment. The corrected split-half procedure would be useful for a wide range of tests since it can be used with various item formats and scoring schemes. However, care must be exercised so that methods of forming test splits do not distort reliability estimates (e.g., including all items of one format in the first half and items of another format in the second would tend to underestimate total score reliability, whereas an odd-even split on a highly speeded test would overestimate reliability). In situations where assessments are clearly criterion-referenced, with a well-defined cutoff point, it is possible to use the $p_0$ or $\kappa$ statistic.

Inter-observer or inter-rater correlations should be computed in situations where several teachers in common score certain student behaviours or products, such as a set of essays. These correlations would indicate the level of teacher agreement, which is an upper-bound estimate of the reliability of scores awarded by individual teachers. Further estimates of score reliability could be obtained by correlating subjectively determined scores with those obtained by another method, assuming that the same skills are being assessed. Computing inter-observer correlation is possible, and at least provides teachers with information on their scoring practices. This is particularly important since much important learning can only be assessed using subjective procedures (see point 5 below).

Typically, teacher-made high-stakes assessments are used to obtain single scores, which are then combined across assessments to form a composite score for interpretation as student grades. For these assessments reliability of composite scores should be computed, and procedures exist for the estimation of composite score reliability (e.g., Feldt & Brennan, 1989). The generalized Spearman-Brown formula can be used to predict the reliability of the composite from the average of estimates of the reliability of individual assessments, provided the weights given to individual assessments do not vary too greatly. Again, the nature and use of the assessment should dictate the method of computing reliability of individual assessments.

It is desirable for teachers to compute reliability estimates for selected high-stakes assessments, but it is unlikely that they will do so unless school-based supports are available that encourage this practice and provide assistance in both computation and interpretation. Clearly, there is no one criterion for the level of reliability that assessment procedures should obtain. The criterion is dependent on the purpose of the assessment(s). It is also dependent on the sources of error which are taken into consideration (e.g., item format, length of time between assessments). There are also some ceilings that could come into play with certain assessment formats (e.g., it is extremely difficult to obtain interrater correlations much above .80 in the scoring of student essays). Frisbie (1988) does venture some suggested levels of reliability, for example, .85 if the scores are to be the only information used in decisions regarding an individual, and .65 for group decisions.

The point of the exercise, though, is to identify the possible error in a given measurement or composite score. Classical test theory provides a simple formula for estimating this from the reliability and standard deviation of the scores, although it is questionable whether the standard error of measure is equivalent throughout the score range. The interpretation of the standard error requires understanding of descriptive statistics and normal distribution, and this could not be expected of teachers at present since there is evidence that they neither use nor understand standard deviation (e.g., Gullickson & Ellwein, 1985; Newman & Stallings, 1982; Webster, 1987). What teachers ought to understand is that there is always error about any measurement, and they should have some idea of what a probable range of this error might be (e.g., 90% confidence interval) for given levels of reliability.

5. Assessments that involve subjective judgments on the part of teachers are an important, integral part of teaching. Authentic assessment often involves observing students applying their skills in actual situations (Wiggins, 1989a, 1989b), and therefore good observation techniques must be used. Even simulated situations, such as those designed for the classroom or the laboratory, often require direct observation and subjective judgement on the part of the teacher. Most experienced teachers understand that subjective forms of assessment are prone to unreliability and to personal bias, but they must know practical procedures which can ameliourate these problems.

Observation is very important as a basis for assessment, particularly at lower grade levels (e.g., Dorr-Bremme & Herman, 1986; Salmon-Cox, 1981; Stiggins, Conklin, & Bridgeford, 1986; Webster, 1987). However, there is little evidence that observations for assessment purposes are conducted systematically. In fact, Stiggins and Bridgeford (1985) report that teachers make substantial use of what the authors call spontaneous performance assessment. This is direct assessment of students' performance that is not designed in advance of the observations and that is not carefully structured. It is important that at least some assessments that are based on observation (particularly those which have high-stakes implications) should be conducted more systematically. This can be done in many ways, such as by identifying clearly what is to be observed, setting forth systematic schedules for the observations (e.g., time intervals with students randomly scheduled to be observed), conducting more than one observation on a given student, carefully recording observations during or immediately following the observations, and using multiple observers (this can include students observing one another as well as other teachers). In all cases records should be maintained of the observations and subjective judgements, and these should be as direct and immediate as possible to minimize the effects of time delay (e.g., forgetting, interfering events). These records provide the evidence for evaluative statements and summative accounts. They also permit checking the consistency of the observations.

## Validity

The validation of assessment is a process that attempts to answer the question, Are the inferences and actions based on the assessment correct? In Messick's (1989a) terms, this becomes: "To what degree--if at all--on the basis of evidence and rationales, should the test scores be interpreted and used in the manner proposed?" (p. 5).

This notion of validity includes both the basis of the evaluation, what is to be considered in the measuring process, and the purpose of the evaluation, what judgement or decision is to be made. Even so, these aspects are only part of what Messick (1981, 1989b) considers to be the heart of validity, construct validity. Construct validation is the attempt to ground the construct presumed to underlie the measurements in a *nomological network* (Cronbach & Meehl, 1955). Messick (1975) makes the point that even for practical testing purposes empirical (criterion) validity and logical (content) validity are not enough; they are but two aspects of the more general need to identify theory and evidence in support of the assessment device as well as the social implications of the intended evaluation. He states that "the meaning of the measure must also be pondered in order to evaluate responsibly the possible consequences of the proposed use" (Messick, 1975, p. 956). This is the unified view of validity taken in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1985), in which it is clearly noted that validation should be broadly conceived and evidence of several types ought to be provided in support of validity claims. The three categories of construct-, content-, and criterion-related validation are primarily for convenience. Recently Anastasi (1986), Angoff (1988), Cronbach (1988), and Messick (1989a, 1989b) all reaffirm the primacy of construct validity for assessment, and indeed for all observational procedures, for any interpretative inference.

Cronbach (1988) prefers to think of validity argument, rather than validation. This is a never-ending process of public study and debate, "validation speaks to a diverse and potentially critical audience; therefore, *the argument must link concepts, evidence, social and personal consequences, and values* [italics his]" (Cronbach, 1988, p. 4). He describes it colorfully as "a public spectacle combining the attractions of chess and mud wrestling" (p. 3). Messick has on a number of occasions (1975--note quote above, 1981, 1988, 1989a, 1989b) identified validity as a question of value as well as one of meaning: "validity is an inductive summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use" (1989a, p. 5). He outlines four basic questions that identify how validation of an assessment can be addressed:

1. What balance of evidence supports the interpretation or meaning of scores;
2. What evidence undergirds the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting;
3. What rationales make credible the value implications of the score interpretation and any associated implications for action; and
4. What evidence and arguments signify the functional worth of the testing in terms of its intended and unintended consequences. (Messick, 1989a, p. 5)

The concept of validity and how it applies to classroom assessment are further discussed under the headings recently endorsed by Messick (1989b), but originated by Loevinger (1957): substantive, structural, and external components of construct validity. These stress the significance of the construct and imply that there are various sources of evidential support. Issues of the assessment's meaning, test authenticity, and content appropriateness are discussed under the substantive heading. The methods for obtaining scores from assessment procedures are considered under structural validity, although in this paper score referencing, such as criterion- or norm-referencing, is discussed in the section called Utility. Test bias, fairness, and equity are discussed in conjunction with the relationship of the assessment to other assessments under the heading, external validity. The four questions posed above are addressed under the three headings as they are seen to apply.

**Substantive component of validity.** Messick (1989b) identifies two major aspects of the substantive component, the first being the familiar notion of content validity: that is, the relevance and representativeness to the content domain of the assessment tasks or items. The second is empirical in nature, and confirms the assessment tasks or items on the basis of data, such as factor loadings, item homogeneity, or known-groups differentiation. Evidence that supports the meaning of assessment scores would be considered substantive.

Are the evaluations content-valid? Do they lead to valid decisions about student achievement? This is the approach to validity typically recommended in the test construction literature for achievement testing (e.g., Crocker & Algina, 1986; Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Nitko, 1983). Validity in the classroom assessment context is usually conceived of as being built into the test development process rather than established on the basis of theoretical considerations and empirical evidence. The process of validating achievement tests has been primarily a task in which judgements are made by knowledgeable individuals regarding the ability of the test and test items to reflect the domain of interest. As such, it is often considered to be different from construct and criterion-related validity. For example, Tuckman (1975) reserves the label of validity for the latter two approaches and applies the term appropriateness to questions of test content. Indeed, Messick (1989b) states that if content validity is construed in this narrow sense, "so-called content validity does not qualify as validity at all" (p. 17). Ebel (1983) has claimed that if there is an explicit rationale for the test, such as a clear operational definition of the assessment domain, this is sufficient for determining the test's content validity. The distinction made by Roid and Haladyna (1982) reflects this position:

> The methods for generating test items may be used in either of two contexts: operational definition or construct validity. The former leads to test scores that are interpreted solely on the basis of what the items broadly represent. That is, there is common agreement that the item and the trait are matched. The latter is based on the underlying meaning of test items as posited by a theory of the relations between a test and other variables in the real world (a nomological network). (p. 8)

However, Messick (1975, 1989a, 1989b), as well as others such as Cronbach, argue convincingly that an appeal to meaning in test score interpretation always requires validation, which implies the need for evidence in support of an underlying construct, this evidence being both logical and empirical.

The judgmental process in substantive validation comes suspiciously close to face validity. This is not to deny the value of the judgmental process and of face validity in its own right, as argued by Nevo (1985). In fact, he claimed that face validity is important to educational assessment. The assessment should appear appropriate to the content and purpose of the assessment both to students and to others, such as parents, who would have an interest in the assessments being valid. Wiggins (1989b) also notes the importance of face validity in that to support learning, assessments must reflect the important outcomes of education and must be publicly perceived to do so. In this way, face validity addresses one of the consequential implications of assessment (Messick, 1989b). But it is necessary also to highlight the need for more extensive evidential and logical support for claims of content validity, if only because of the fallibility of judgements. In part, too, the typical judgement approach to validity often misses a crucial point--that validity is a characteristic of test scores and test score use, not of the test itself, and must be linked clearly to the purpose of the assessment. This point is made by many writers, including those who are attempting to address teachers (e.g., Crocker & Algina, 1986; Gronlund & Linn, 1990), and is clearly enunciated in the *Standards* (American Educational Research Association et al., 1985). Content validation should be built into the procedures for assessment development by providing for appropriate expert review and possible empirical support (general procedures such as those recommended by Hambleton, 1984). Persons involved in content validation should be clear as to the intended use of the assessment and the context in which it will be applied. In schools it may not be practical or even reasonable to have expert review of classroom assessments, but it is possible and desirable to have a substantial sample of the most consequential assessments used by a teachers reviewed by other teachers, curriculum specialists, and supervisory personnel, and even available for public review. This is the process that preceded the Chambers and Fleming studies in Cleveland (e.g., Chambers, 1982; Chambers & Fleming, 1982).

It is not enough to determine if the items appear to assess the domain of interest, they ought also to have instructional fidelity. Nitko (1989) states it this way: "Prescriptions for test designs . . . are obviously linked to the instructional methods to be employed, to the instructional conditions under which instruction is to occur, and to the instructional outcomes to be expected" (p. 448). This emphasizes that the assessment tasks must correspond to the instructional activities designed by the teacher, but it also includes consideration of the in-class environment in which it takes place. For example, in a science class where emphasis was placed on designing and carrying out experiments in a laboratory, it is of little value, and probably counterproductive, to use a test that asks students to list the steps in an experiment or to give a definition for the term hypothesis, even though it could be argued that these items would fit a related domain. The assessment ought to be based on the most important science processes and content for which the instruction was intended and designed. It also ought to reflect the means by which the students learn the material, and the setting in which this kind of learning is to be displayed. In short, assessment in the classroom should both reflect and support the learning. This is what Wiggins (1989a, 1989b) means by ***authentic*** assessment: "We need to observe students' *repertoires*, not rote catechisms coughed up in response to pat questions" (1989b, p. 706). Wiggins (1989b) goes on to emphasize the need to identify the most important learnings for students, and what students should be expected to do in various discipline areas (the performances or standards). This would include such tasks as "write a brief history of your family", or "conduct research to determine the effects of caffeine on activity" rather than answer questions about these topics. The actual student behaviours and products would form the focus of the assessment, which would result in applied performance assessment at its best (as described by Stiggins, 1987b). These performances and work samples would be accumulated in portfolios or records of achievement, such as in the procedures currently being developed by the Assessment of Performance Unit (Burstall, 1986), and others, in Great Britain (e.g., Fairbairn, 1988). These approaches to evaluation have long been advocated in areas of the arts and in the teaching of writing (e.g., Wolf, 1989), and are coming to the fore in a variety of curricular areas in large-scale testing (e.g., Mullis, 1992) as well as in classroom assessment (e.g., Wolf, Bixby, Glenn, & Gardner, 1991). Indeed, as noted by Haney and Madaus (1989), there is nothing really new here accept to make them more usable in the classroom, and more systematic and valid.

What is possible in the classroom setting is that teachers can clearly outline the purpose(s) of their assessments, the focus and context of the assessments, and the procedures they used to develop the assessments so that these can be reviewed by their students, colleagues, and administrators, and by parents and other members of the public. The constructs that teachers attempt to assess are not formalized in the scientific sense, but they are described in terms of achievement as it is known in a particular grade level and subject area. Most classroom assessments are intended to provide evidence of the learning of content by students; this is particularly so at the post-primary school level (the case studies; Dorr-Bremme & Herman, 1986; R. J. Wilson, 1990). The instructional objectives are generally outlined by curriculum guides and the details of content are reflected in the learning materials used in the classroom (primarily textual). As such, then, the major substantive validity question is one of content validity. The purpose and context of the test should guide any judgmental review of its content and format. Probably those teachers and administrators who teach in similar settings should be the judge of the substantive validity of an assessment.

Empirical arguments can be brought to bear on the substantive validity of classroom assessments. Item analysis is one procedure that is practical in the classroom and that can be used to determine if particular items perform as they should relative to others, but it is well known that teachers do not use the numerical methods of item analysis (e.g., the case studies; Gullickson & Ellwein, 1985). However, simple comparisons of item results can be obtained by teachers to see if items that assess similar objectives tend to operate similarly; that is, "to appraise their homogeneity as diagnostic categories" (Messick, 1989b, p. 68). This is necessary if the tests are to be used for informing instructional decisions, which they often are. Homogeneity is a necessary condition for meaningful combining of scores based on several items, or information from several sources, purporting to assess the same concept It can be thought of as answering the question, do these measures all provide evidence for assessment of the same "substance". In the case studies teachers generally looked at student performance on individual items or on sections of the tests, and item analysis would make the process more systematic and defensible.

**Structural component of validity**. Structural validation is another approach to internal validation of the assessment. It includes the "fidelity of the scoring model to the structural characteristics of the construct's nontest manifestations and the degree of interitem structure" (Messick, 1989b, p. 43). This is related to the homogeneity aspect of substantive validation, but extends it to how the scores are obtained: total score, profile of subtest scores, or some other scoring method, or a qualitative evaluative statement or description. Clearly, the purpose of the assessment and the content being assessed must be considered to identify the structure of the assessment and the appropriate scoring and evaluation procedures.

For most classroom assessments teachers obtain the students' total scores on the assessment by summing the marks obtained for all the items or tasks--this Messick (1989b) describes as the "cumulative quantitative model" (p. 44). This assumes an additive, compensatory model of assessment tasks or test items. That is, the marks for items can be added (the trait is sufficiently homogeneous and unidimensional), and the scores of two students that are numerically equal but obtained by getting different items correct are identical. These two assumptions are indeed very questionable in classroom assessment (as they are for many, if not most, assessments of human behaviours; e.g., Traub & Wolfe, 1981)! For example, two items may represent equally important concepts and be of equivalent difficulty for students generally but may well tap different skills. The summed scores for two student on these two items are the same if one student knows only one concept and the other student only the second concept, yielding a score of "1" on these two items if the correct answer was awarded one mark and one answer was correct and the other incorrect. The score does not mean that the student did not have either requisite skill. The item scores simply may not be additive. Furthermore, one item on a test may require considerably greater understanding than another, yet be awarded equivalent marks; these two items are not compensatory. Two students of differing abilities could both get the simpler item correct and the more difficult item incorrect, suggesting that they have the same skill levels. There are solutions to these problems, at least in theory, using item response models, but these are not practical in classroom settings.

The previous paragraph poses a conundrum from which it is difficult to escape. Domain-referenced assessment is an attempt to avoid the problem by making the domain so narrow and well-defined that assessments can be developed from it that do not lead to ambiguous interpretations of scores. One such approach is that of Roid and Haladyna (1982), who, among others (e.g., Bormuth, 1970), propose the use of language transformations and item forms to produce well-defined test domains. However, although there is evidence that teachers can learn these procedures (Roid & Haladyna, 1982), they are not possible to use on an every-day basis in the classroom because of the diversity of subject material and contexts which teachers encounter; recently this has been clearly recognized by Popham (1992). It is extremely difficult to specify domains this precisely, nor is it readily possible to design assessments that have the required purity. Furthermore, this often leads to a plethora of *itty-bitty* skills and tests, which flies in the face of the integrative learning espoused by many pedagogues (e.g., Haney & Madaus, 1989; McCormack & Yager, 1989; Nickerson, 1989a Wiggins, 1989b; Wolf, Bixby, Glenn, & Gardner, 1991). Teachers apparently understand the impracticality of domain-specific tests. They do not even seem to have much interest in materials based on this model, such as Popham's IOX materials. As noted earlier, the primary purposes of classroom assessment are grading and informing instruction. It is extremely difficult to assess every skill in a detailed way, and, as well, to form some composite score based on the accumulation of skills. Aggregate scores and grades have a problem of ambiguity when formed from an accumulation discrete skills (e.g., Stiggins, Frisbie, & Griswold, 1989; Terwilliger, 1977, 1989).

It appears that classroom assessment generally will have inherent ambiguities in interpretation, except in the rare cases where assessments are designed for individual diagnosis. Perhaps it is only possible in the classroom to differentiate assessments for conflicting purposes, such as cumulative grading and instructional diagnosis. It is possible to design formative assessments which attempt to assess some well-defined domains. In these cases profiles of scores can be obtained. For summative purposes, including grading and overall reporting, assessments will continue to include multiple skills. The problem here is to ensure that the most important skills receive the highest weighting and that this is not offset by numbers of items and scoring practices (e.g., Oosterhof, 1987; Terwilliger, 1989), or by some other artifact of the measurements and the aggregating equation (Thayer, 1991).

An alternative assessment practice is to set tasks that require cumulative skills and integrated understanding for successful completion. Evaluation of these cumulative skills is usually subjective. This is the approach recommended by educators like Wiggins (1989b). An additional approach is to accumulate assessment information throughout a series of meaningful tasks, and to use these as a basis for evaluation (e.g., work portfolios). Some elementary schools have opted for narrative reporting of student progress. In these settings, the focus of evaluation is on the in-class activities and performances of students. This results in considerable reliance on teacher judgements of students' ongoing work in the determination of achievement. This type of evaluation suggests the use of records of achievement or work portfolios.

**External component of validity.** The external component is far-reaching and complex since it encompasses all approaches to validation that make use of measures external to the one being validated. It involves substantive theory to identify meaningful relationships to other variables, the nomological network, and criterion-related validity to verify testable predictions. It is exemplified by Campbell and Fiske's (1959) multitrait-multimethod validation, which is one of the more stringent approaches to construct validation. This suggests that the measure of a trait must have greater correspondence with alternate measures of the same trait using different methods than with different traits using similar measuring techniques. Finally, it also seeks to determine if the measure is related to treatments or group differences to which it should be related, and independent of population subgroups from which it ought to be independent, such that it is equally appropriate to these subgroups. Ascertaining test fairness is an obvious significant part of external validity. But what does all this imply for classroom assessment?

In the classroom, there is rarely the opportunity to study in detail the constructs that might underlie various assessments, so that a theoretical network is unlikely to be determined. Failing a formal network of theory and measurement, it is possible to employ some of the techniques of external validation to classroom assessments. For example, Gronlund (1985) notes that assessments of achievement should be sensitive to good instruction on pertinent material, so pretest-posttest differences should be large on assessments related to instruction. This principle is almost too obvious, but it is rare in every day practice for teachers to conduct pretest-instruction-posttest checks to validate the test (e.g., the case studies). Occasionally pretest-posttest procedures are used, but to determine if learning takes place rather than to validate the assessment. Teachers would also view using these procedures to validate the assessment as impinging on valuable and limited instructional time. However, teachers do use informal ways of checking their assessments. They often check to see how their "better" students did on the assessment, under the assumption that the performance of these students should be well above average. Better students would have been identified by earlier assessments, making this a very rough-and-ready validity check (it may also suggest a halo effect if subjective forms of assessment are used).

It is simply not practicable for teachers to provide systematic, external justification for every assessment they conduct. Furthermore, for many of the variables that are the focus of classroom assessments theoretical developments are not at the stage where broadly based validity claims can be made (e.g., for particular objectives of the curriculum very little validity research may have been done, and the nature of the relationships among objectives is not clear). Of course this presumes that the objectives and content are imbedded in a larger curricular design, and that issues of construct validity are of lesser direct relevance to the teacher.

The final aspects of validity to be considered here are those of the value implications of score interpretations, and the personal and social consequences of decisions based on assessments. Part of the reality of assessment is that the measures obtained on students are given labels and meaning, such as achievement or learning. This does not mean that it is unacceptable to speak in value-laden terms, but this does suggest that care must be taken to ensure that assessments relate to important aspects of learning and that they are appropriate to the students. Negative implications for students must be recognized.

An aspect of external validation within the realm of classroom assessment is fairness in test use. The questions to ask here are: Are the evaluations fair to the students? Are they free of particular teacher biases and prejudices that may impact negatively on students? Do they serve various groups of students equally? Fairness is typically considered to be freedom from bias against particular persons for reasons other than the trait being measured, freedom from "prejudice or having a particular bent or direction. . . . or unfair to groups or individuals characterized as different from the majority of test takers" (Tittle, 1988, p. 392). Bias in assessment is clearly a problem of validity since biased assessments yield invalid results for particular individuals or groups (S. B. Anderson, Ball, Murphy, et al., 1975; Berk, 1982a; Cole & Moss, 1989), but bias reflects a particular kind of invalidity, one resulting from an instrument assessing different processes in different groups such that one group is favored over another (Berk, 1982a). This notion of bias is different from the narrower one of rater or observer bias, such as leniency/severity error, halo effect, and logical error (e.g., Gronlund, 1985; Anastasi, 1988). For this paper fairness is considered to be the opposite of bias, freedom from bias, and no particular individual or group has an opportunity to perform better, or the likelihood of performing poorer, on the assessment for reasons other than the construct being assessed, given that the student affords a real effort.

Bias in testing, as the expression is used today, is ambiguous (Cole & Moss, 1989). Sometimes it is restricted to the internal characteristics of the test (e.g., item bias) and sometimes it refers to the use of the test itself (i.e., fairness of test use). This distinction is important to make since the two notions lead to different methods for studying bias. Judgmental procedures are commonly used to check for internal characteristics that may be biased, answering questions such as, Is it appropriate and fair to test for the intended trait in this way? Empirical procedures are used to check fairness of test applications, Is the predictive validity different for different groups?

Fairness in testing is part of external validity, but also includes concerns of social and moral appropriateness. Cole and Moss (1989) describe the values surrounding the purposes, intended and unintended uses, of an assessment as being outside their concept of validity, but clearly note the importance of social justice concerns. They restrict the notion of bias in construct validation to test use and to "differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (Cole & Moss, 1989, p. 205). However, for purposes of this paper the broader notion of bias will be retained and subsumed under the external component of validation.

It is important for classroom assessment that both conceptualizations of bias and fairness be considered. The first one is important in the design and development of the assessment procedures and instruments. This involves the control and explication of personal preferences, and the removal of personal biases and prejudices insofar as they may reflect negatively on known subgroups of the population (reflected in such things as the choice of nonsexist language, removal of negative ethnic stereotypes, etc.). It should be noted that this type of bias may not have a direct biasing effect on performance of subgroups on a particular assessment, but it may do so in the longer term, such as by causing negative reflections on a subgroup.

The second comes into effect when measurements or observations are interpreted and used. Here, personal subjectivity must be understood and controlled so that the interpretations reflect as fairly as possible the true performances of the students. It would include such things as how assessments are weighted to obtain summative scores and grades.

A notion related to fairness is that of objectivity, or freedom from individual subjectivity. In general, there is no such thing as objectivity in assessment. Subjectivity enters into the intent of the assessment, the student characteristics to be assessed as well as the indicators of these, the scoring procedures, and the value interpretations of the scores. However, the narrower concept of objectivity in scoring student responses is discussed in the section on Utility below. This is where rater or observer bias, such as leniency/severity error, halo effect, and logical error are considered.

There are procedures which can improve the fairness of classroom assessments. Judgmental procedures are necessary to reduce the possibility of internal bias in assessment procedures and empirical procedures are necessary to eliminate unforeseen bias effects. These are presented in the section below, Recommendations for Validation.

Classroom assessments are so frequent and so pervasive that care must be taken that the assessments do not become vehicles that permit or perpetuate inequities and prejudices that may be commonplace in society. It is incumbent upon teachers to ensure that personal preferences are clear and explicit so that they are apparent to students and the public, as well as to the teachers themselves and their colleagues. Teachers must be particularly sensitive to subgroups in their classrooms so as not to advantage or disadvantage them in ways that are unfair.

**Recommendations for validation of classroom assessments**. There are seven recommendations regarding validity for classroom assessment practices and for the preparation of teachers. Recommendations 6 and 7 refer to the logical or content aspect of substantive validation, recommendation 8 refers to the ability of the assessment to provide the information for particular instruction-related decisions, and recommendation 9 refers to the empirical aspect. Recommendation 10 relates to the structural component of validity. Recommendations 11 and 12 relate to external validation, and of these recommendation 12 deals with fairness and equity.

Classroom assessment must reflect **that which is most important** to learn in terms of the content discipline or the field of study (e.g., knowledge and skills that are accepted as fundamental to thinking in the discipline), and in terms of important life skills (e.g., higher level thinking skills and skills that lead to transfer of learning). This is part of the social responsibility aspect of validity. Classroom assessment must also reflect **the way in which learning occurs** in everyday life and in a field of study, such as problem solving of practical, real-life problems using the knowledge and processes developed in a subject discipline. The format of the assessment should be similar to the particular classroom and other activities which gave rise to the learning, fidelity of the assessment. All of this is what Wiggins (1989b) means by authentic assessment, and Cole (1987) and Nitko (1989) mean by integration of instruction and assessment.

The validity of classroom assessments can go awry in the specification of the behaviors which are to represent the domain, in the sampling of behaviors to be assessed, in the designing of assessment procedures (test items, etc.), in the weighting of items and marking of student responses, in the combining of marks to form student scores, and in the interpretation of student scores to form evaluative judgements. These are dealt with below, although the problems of interpretation are discussed in the Utility section.

**6.** Classroom assessment procedures must provide information on the learning that is most important for our students, including complex and higher-level thinking. Therefore, teachers must have the knowledge and skills to design and conduct classroom assessments that are most appropriate to these kinds of outcomes. This includes both the content and thinking processes identified in subject areas, and the broader thinking skills, communications skills, and cognitive strategies (metacognition) that are currently emphasized by our schools.

Today, the important concepts and skills are clearly identified in various subject areas. Almost any provincial curriculum guide spells these out in considerable detail, and they are continually emphasized in the writings of recognized curriculum authorities (e.g., Klopfer, 1971, and Yager, 1987, in science; National Council of Teachers of Mathematics, 1989, in mathematics). The subject educator specialists are the most important source of the "what" of classroom assessment. They and the curriculum documents, teaching materials, and related professional journals clearly identify what is relevant to assess in particular programs and at various grade levels.

There is also ample writing today from a psychological perspective on the importance of higher-order thinking skills to classroom learning. Although the paucity of procedures for assessing higher-order thinking skills is bemoaned (e.g., Nickerson, 1989; Wolf, Bixby, Glenn, & Gardner, 1991), there is considerable effort going into the clarification and development of procedures for assessing these kinds of skills (e.g., Baxter, Shavelson, Goldman, & Pine, 1992; Norris, 1989; Norris & Ennis, 1991; Stiggins, Rubel, & Quellmalz, 1988). There are a number of models, including the classic Bloom's taxonomy (see Bloom, Hastings, & Madaus, 1971) and those of others such as Biggs and Collis (1982) and Stiggins, Rubel, and Quellmalz (1988), that can be used to help guide assessment in this regard.

There is considerable evidence that the assessment procedures commonly used by teachers do not tap higher-level thinking to any great degree (e.g., Fleming & Chambers, 1983; Haertel, 1986; Stiggins, 1986; Stiggins, Griswold, & Wikelund, 1989; the case studies). There is also evidence that teachers have difficulty producing assessment questions which require higher-level thinking of students (e.g., Carter, 1984). This is the true craft of classroom assessment, and should not be short shrifted in teacher preparation. Teacher preparation should include the opportunity to produce a variety of assessment devices that focus on important, higher-level cognitive learnings. There is a vast array of materials, including many excellent textbooks (e.g., Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Nitko, 1983; Popham, 1990), that can be used to show prospective teachers how to construct quality assessment instruments. There is also considerable material which provides guidance in preparing assessments pertinent to particular subject areas (e.g., Bloom, Hastings, & Madaus, 1971). Both teacher training and professional support in the field should assist teachers in preparing the kinds of assessment necessary for higher level thinking; workshop and inservice materials are now readily available (e.g., Stiggins, Rubel, & Quellmalz, 1988).

**7.** Classroom assessment must reflect good pedagogy, both to support the learning process in the classroom as a mode of mental development by emphasizing what is important and setting standards for its attainment, and to provide practice for this kind of thinking (e.g., thinking skills that integrate learnings from several content topics or areas, and that impinge on future learning). Teachers must have training and actual experience in preparing and using assessment procedures that support the most important learning in classrooms, those that should be emphasized in class and on the assessments. Teachers must know how to develop and use a wide variety of assessment procedures. These include paper-and-pencil techniques, but more importantly, they should include systematic observation and applied performance assessment.

Good pedagogy suggests that contextually-based learning is far more powerful than strictly lecture and question-answer formats or text-based instruction. Students must have the opportunity to engage in tasks and problems of real significance both to them and to the world outside the classroom. Without having the opportunity to "try on" their learning in many contexts, and in significant ways, the learning remains isolated, often irrelevant, and nonfacilitative to future learning or work. The integrative and implicative aspects of learning and knowledge must be fostered in teaching and demanded in assessment (Wiggins, 1989a, 1989b). The effects of evaluation on what is learned and how this is learned are well recognized (e.g., Crooks, 1988; Fuchs, Deno, & Mirkin, 1984; LaMahieu, 1984; Smith, 1991; Smith & Rottenberg, 1991).

The form that assessment takes is almost as important as the function and content of it. Significant learnings must be assessed in ways appropriate to the representation of the skills. This reinforces what is important in learning and clarifies how it must be exemplified, thereby enhancing both instructional fidelity and face validity. The social consequences are apparent: learning must be exhibited in meaningful ways, and judgements based on assessments should therefore reflect the important abilities of students and how, and in what settings, they can apply these skills.

The quality and level of language used in an assessment is important. The language that is appropriate to the content being assessed should be used, provided it does not detract from the students' ability to respond to the assessment given that they understand the most important concepts. The language and directions that instruct the student in the assessment must be as simple as possible, and must be clearly understood so that students answer the correct question, and so that aspects of other skills that are not the focus of the assessment do not detract from student performance. Both of these issues are apparent in Yager's (1987) plea to "Test science not reading". This does not deny the importance of correct and precise language in understanding and communicating important ideas, but care must be taken that language does not get in the way of the students' ability to demonstrate their understanding of the concepts and skills for which the assessment is intended. This too is true professional craft, and prospective teachers must have opportunity to practice it prior to entering the classroom.

**8.** Classroom assessment must be designed in such a way that it can provide the correct information for the intended purpose. Teachers must have the training to devise assessment procedures for a variety of purposes, such as for grading of students, individual and group instructional diagnosis, and evaluation of instruction. There are other purposes, but these are likely to be secondary. Teachers must also know when and how it is possible to accommodate more than one purpose, and to recognize when this leads to incompatible procedures.

Teachers are not well versed in the development of assessment for instructional purposes, formative assessment (e.g., Stiggins & Bridgeford, 1985; the case studies). The appropriate procedures are dependent on the subject matter, the developmental level of students, and other factors, like language background of students. This makes it difficult to use any one assessment procedure, or to use general guidelines for the development of all formative classroom assessments. The areas where substantial amounts of diagnostic assessment materials are available are mathematics and reading, and this is primarily at the lower grades. Therefore, teachers ought to have specific training in the development of formative, instructionally diagnostic, assessment procedures for one or two subject areas and for certain grade levels of students. Teachers should be familiar with the use of subtests, or portions of a test, and be encouraged to use them for instructional purposes. The case studies show that teachers do not typically do this at present, except very informally.

**9.** It is important for classroom assessment that several tasks or items devised to assess the same skill or concept yield similar results with students. A check for this presumed homogeneity is one way of testing if there is a meaningful, stable underlying construct, and that subtest scores can be interpreted. Without this, the construct can have little utility in future learning. Teachers must know how to check for homogeneity. This involves identifying the assessment tasks or items which should "hang together" and checking student results to see if this is the case.

Teachers are not likely to use anything more than the simplest of statistics with their test scores (Gullickson & Ellwein, 1985; Newman & Stallings (1982). But it is important that teachers have some systematic procedures with which to check their item results systematically. This need not be the traditional forms of statistical item analysis, but certainly it is useful for teachers to know such characteristics as the difficulty level of items (proportion of students getting item correct, or item mean). It is also useful for teachers to have a way of checking responses to an item against those for another within a proposed subtest. Crosstabulation may be the simplest procedure, since it could readily incorporate assessment tasks which are scored on more than a two-point scale. Small computers are available in schools today, which make the task less onerous, although programs may be difficult to obtain (spreadsheets can be used for this, but it still involves a fair amount of work). One alternative would be for teachers to identify a few students who did well on the assessment and a few who did poorly, and to compare their results across the items which are designated as belonging to a particular construct.

**10.** Classroom assessments must be designed to provide the scores and information for which they are intended and these scores must be appropriate reflections of the importance of the content. Teachers must be trained to obtain scores from assessments corresponding to the nature and purpose of the assessment. This involves identifying clearly in advance of producing the assessment, the approximate importance, and hence the weighting, of various areas of content and levels of skills (essentially producing a table of specifications for the assessment) and maintaining this intended weighting in the actual composite scores.

The scoring system and procedures must be carefully designed to produce results that are appropriate to the purpose and focus of the assessment. For instance, if diagnosis is the focus, then there must be parts of the assessment that are directly relevant to each diagnostic category (construct meaningfulness) and they must be distinguishable from one another (interpretability). The level of diagnosis (e.g., individual versus group) would dictate the specificity of categories and the reliability requirements for assessment in each category (consequential significance).

In summative assessment where a composite score is obtained, the scoring procedures must retain the relative importance attributed to various aspects of content and process. This should not be distorted by numbers of questions, difficulty of question or task, or relative weighting. A test blueprint or table of specifications would assist in clarifying the content, levels and kinds of thinking, and relative importance of various components. Teachers do not typically use these procedures for producing tests (the case studies), but often do give students a clear indication of the content that is to be assessed and the kinds of test questions that will be used. Test specifications should be identified explicitly prior to the assessment and should be made apparent to students so that students know where to put their effort.

There is evidence that teachers do not provide careful directions to students on their tests, or indicate clearly the marks awarded (e.g., Fleming & Chambers, 1983; the case studies). This may distort students' efforts relative to what is deemed most important for the purposes of the assessment. Depending on the nature of the content and the purpose of the assessment alternative scoring schemes may be devised. In some summative situations it may be useful to obtain subtest scores, each of which may be interpreted separately. For example, this would be the case where it is intended that students master the skill assessed by each subtest.

**11.** Classroom assessments must be designed to permit confirmation or disconfirmation of findings from previous assessments or from external sources. Previous assessments or other relevant evaluative information provide a check on assessment results, thereby giving external evidence for their validity. Teachers must base their judgements of students on information from multiple, diverse sources. The validity of the information should also be checked periodically in this way. This can be done informally by comparing the results of particular students on one assessment to those on another. It is preferable to use something more systematic, such as correlation, but it is unlikely that this will be done in the classroom setting. Teachers should understand the importance of external validation. They should attempt to obtain several assessments of students' learning using different assessment procedures, and check for glaring discrepancies.

Multiple assessments of important content and concepts lend credence to the score interpretations and the resulting value judgements and actions. This is a reliability concern if the assessments are virtually identical. But more importantly, if the procedures for assessments differ but are of similar good quality, it allows students several opportunities to exhibit their learning, bolstering the credibility of any resulting judgements and enhancing their fairness. Teachers are clearly aware of the need for gathering multiple forms of data for major decisions, as evidenced by the assessment information they include to determine grades for reporting (e.g., Dorr-Bremme & Herman, 1986; the case studies).

The methods for checking the results of one assessment against those of another vary depending on the nature of the assessment and the external criterion, although various types of correlation are the most common in educational measurement. Teachers cannot be expected to calculate correlations between sets of scores on assessments, but modern computing equipment should make this facility available in schools. It may be reasonable to have teachers conduct checks between the results of various high-stakes assessments, particularly at the higher grade levels where teacher-made or school-level examinations are commonplace.

**12.** Classroom assessments should be free of biases and prejudices that may have negative impact on students, particularly as these might affect known subgroups of the population (e.g., gender and ethnic groups). There are two kinds of bias in assessment, one of which may have direct deleterious effects on the assessment results of particular groups of students by virtue of their differences in experience that make them perform below their actual skill level. This is the issue of fairness or equity. The other is more subtle, and may not have a direct effect on results, but it is bias or prejudice that has a negative reflection on particular subgroups and may over the long run have a negative impact on these subgroups. Teachers must be trained and experienced in recognizing biases of both types in the materials they use, in the expectations they have of students, and in how they relate to students. There are judgmental and empirical procedures that can be invoked to reduce bias.

Judgements can be made by knowledgeable colleagues, but there is also the need to submit classroom assessments to public scrutiny. Cole and Moss (1989) suggest that it may be useful to include representatives of identified subgroups in the reviewing process. The judgmental procedures are applicable to internal bias in assessments (such as item bias), and apply to both the fairness of assessment and its freedom from prejudiced portrayal. The following steps are based on Cole and Moss (1989), Shepard (1982), and Tittle (1982):

1. Logical analysis to establish the relevance of the assessment (items, etc.) to the trait of interest and to the purpose of the assessment, including a clear indication of how relevance was established. The context of the assessment must be clearly considered in this analysis.

2. Review of the assessment to determine its fairness to the intended students, with particular concern for language, stereotypic representations, and material with content or examples that may operate differentially on, reflect negatively on, or be offensive to, subgroups based on characteristics such as gender, ethnic minorities, cultural differences, socioeconomic status, and geographic location. The content of the assessment and the language and material used should be considered. There are suggestions that test content should contain a balanced representation of disadvantaged subgroups.

3. Review of the assessment procedures to ensure that the format and style does not affect subgroups differentially (e.g., use of multiple-choice item types with poor readers when the intent is to assess understanding of scientific principles). There may be guidelines that apply to specific subgroups, such as those whose first language is not the language of instruction.

4. Care in the administration of assessments to ensure that all subgroups are equally cognizant of how to respond. In scoring constructed responses and making observations irrelevant influences must be minimized. Possible influences include different language uses by various subgroups (e.g., the use of the vernacular or of less formal rhetorical patterns) or behavioral styles (e.g., the reticence of various ethnic groups to respond actively in school settings).

There are many empirical procedures proposed to detect bias in test use. These are based on various conceptions of bias identified by Shepard (1982) and Cole and Moss (1989). Many of the procedures involve sophisticated statistical analyses (e.g., factor structure comparisons, item-characteristic curve comparisons), and require large numbers of test scores. These procedures are clearly inappropriate to classroom applications. However, some procedures are practicable in a simplified way for teachers to use. These are (Cole & Moss, 1989, and Tittle, 1982, 1988):

1. Tryout of the assessment with similar students, which can be done in classrooms by trying certain procedures one year to be used for evaluation in subsequent years (this process requires both logical and empirical analyses).

2. Development of the score or data interpretation scheme, which implies distinguishing the observations from evaluations based on them (this can be the setting of norms, but more likely is the application of value-laden descriptors, such as grades, to various scores or observations).

3. Development of assessment procedures that provide additional data to which the interpretations based on the assessment of interest can be compared. This is likely to be too cumbersome for regular classroom use, but what is suggested here is that several different sources of evaluative information be gathered whenever substantial decisions are to be made regarding students.

4. Comparison of item difficulties for various subgroups, with a view to consider the item in particular, and also consider sets of items with differing formats, structures, or procedures.

## Utility

Utility of assessment attempts to answer questions such as: Are the evaluations useful to the student, to the teacher, to the parents, to the educational system? Do the assessments provide information that can be interpreted intelligibly? This is an extension of validity in its consequential sense, but utility emphasizes the effectiveness of the assessment rather than its meaningfulness and legitimacy, and its relationships to other measures and other constructs.

Utility refers to how well the measurement provides information that can be used by those making the evaluations and those directly affected by them. Part of this is obviously the communicability of the assessment information, and the clarity and ease with which it can be interpreted by educators and others. Utility is related to efficiency but also is distinct from it. Certain assessment procedures may be useful for more than one purpose, but may be efficient for one of these purposes and not the other. For example, assessments which provide a wealth of detail with respect to the relevant learnings might be considered high in formative utility to the teacher and likely also to the student, but, although the assessment results could be summarized to provide a readable report, this would be inefficient for providing information to the public.

Tuckman (1975) describes some utility characteristics under the heading of "interpretability" and "usability" but he includes only referencing of test scores and practical aspects of test administration. Gronlund (1985) describes the "usability" of tests as referring to "the *practicality* of the assessment procedure" (p. 57), which includes aspects of efficiency and of utility. The notion of utility that is used here includes the practical characteristics identified by these two authors, except those of efficiency, and defines them in terms of who is to use the information and how it is to be used. Utility includes the referential basis of assessment information (how scores are interpreted), the applicability of assessment results (how the results can be used), and the communicability of the assessment results (do recipients of the assessment results understand the implications of them). The additional aspect of objectivity in scoring is also included.

**Referential basis for interpreting assessment scores**. This refers to the process by which inferences and value-laden statements (e.g., grades, or statements of mastery/nonmastery) are determined from the assessment data. It involves the referencing of scores according to some scheme or procedure, the typical schemes being norm-, criterion-, and self–referenced. These three apply readily to formal testing procedures which provide numerical scores that can be referenced to some scale or point and to which value judgements can be applied. As noted earlier, however, many classroom assessment procedures do not fit clearly into one or another of these referencing schemes. In fact, teachers often use a combination of two or even all three (e.g., R. J. Wilson, 1990; the case studies).

Some procedures may supply evaluative statements directly. Most subjective ratings of behaviors or products do just this (e.g., holistic or primary trait scoring of writing) and in these situations the referencing is not as clear. Even though criteria of performance may be identified, it is difficult to separate the normative influence from subjective judgements. Student responses are usually judged relative to students with whom the teacher is familiar and whose responses give a basis for what is considered reasonable. Rarely is there a clear standard against which the performance and the scores are compared. Many other assessment procedures used in the classroom, for example, work portfolios, records of achievement, and anecdotal records, do not correspond to the referencing categories directly. But evaluative judgements are being made, and the bases of these judgements may well be a combination of two or all three referencing schemes. Furthermore, the primacy of a referencing scheme may vary from one assessment to the next, let alone from one teacher to another.

The call in the seventies and eighties was to bring about criterion-referenced assessment in the classroom (Berk, 1984b, Popham, 1981). In part, this was a reaction to norm-referenced commercial survey tests (e.g., Houts, 1977). The problem for the classroom is not this simple, however. It is possible in certain well-defined content domains (such as adding 2-digit numbers) to specify meaningful criteria of performance, although many criteria that are suggested are purely arbitrary. There is no strong evidence that there is anything sacrosanct about 85% correct for mastery (such as that proposed in Block, 1971), for example. Also, as discussed in the section on validity, these kinds of narrow domains are not that common, particularly at the higher grade levels, nor are they consonant with many important learning outcomes. Teachers have not gravitated to the criterion-referenced approach, either on the basis of a mastery learning philosophy or any other basis. The teachers in Robert Wilson's (1990) study indicated that they made considerable use of criterion-referencing, but they also acknowledged possible norm-referenced interpretations of student results.

It appears that all three referencing approaches have legitimacy in the classroom, but that they are often used in combination. This poses difficulties with the meaning of the evaluative labels derived from the referencing. For example, what does a grade of "A" mean? Does it suggest excellence or a very high level of performance? Is it better than 80% of the class? Does it imply mastery of the content? Does it suggest great improvement in performance? The basis on which interpretations are made varies greatly from setting to setting. Not only do teachers vary among themselves but individual teachers vary from one class to another, year to year, and course to course. There is no single approach used by teachers to form their evaluative judgements, and certainly there is no clear best way to obtain evaluations from assessment data. Although there is only limited research, from recent work it is probably safe to conclude that a wide variety of grading schemes are employed in classrooms, and an even wider variety of reporting procedures (e.g., Friedman & Manley, 1991; MacRury, 1988; Manke & Loyd, 1990, 1991; Stiggins, Frisbie, & Griswold, 1989; the case studies).

There is no agreed upon procedure for establishing standards of performance, and perhaps there ought not to be. Standard setting procedures have been clearly outlined (e.g., Jaeger, 1982; Livingstone & Zieky, 1982) and discussed over the last 20 years (e.g., Berk, 1986; Nitko, 1991), but these tend to be more appropriate to large-scale assessments of student competencies, and they certainly are not without their problems (e.g., Glass, 1978; Shepard, 1984). There is no one correct or appropriate approach for the setting of standards of performance, even for large-scale assessments (Berk, 1986). Some authors, such as Popham (1981) and Shepard (1984), suggest procedures that include review of data on student performance in the making of judgements, while others, such as the Angoff method, rely strictly on analysis of the test questions. It is impossible for classroom teachers to use panels of persons to assist them in setting standards since they cannot be readily established for all the assessments that could require formal interpretation procedures. This does not exclude the possibility, and desirability, of teachers going beyond their own expectations of student performance in setting standards. Certainly other teachers could be involved, and many curricula have implied standards in their statements of goals and objectives. Furthermore, teachers could keep accurate records of previous students' performances on assignments and tests, which would assist them in making evaluative judgements on the performance of future students. This implies that teachers use the same assessments from one group of students to another--which is practicable in some cases. Finally, educators such as Popham (1981) argue that the judgement process is a real and necessary part of evaluation in schools, and emphasize the importance of the context of the assessment (the setting, the students involved, the course) and the purpose of it (its consequential nature) in the setting of standards.

Some schools have explicit policies regarding grading and reporting practices, but the standard expected of students is largely left to individual teachers to set. There are some general guidelines that can advise teachers on how to establish their grading practices, but these are guidelines based mainly on experience and not on theoretical and empirical grounds (e.g., Gronlund, 1985; Stiggins, Frisbie, & Griswold, 1989; Terwilliger, 1989). In fact, there seems to be discrepancy between what measurement specialists would suggest regarding certain aspects of grading and what teachers actually do (e.g., Stiggins, Frisbie, & Griswold, 1989).

The importance of making judgements of the quality of student performance demands that it be addressed in a systematic and defensible way. There is no one procedure that is appropriate in all or even most situations. Perhaps all that can be hoped for is that teachers use informed judgement in setting their outcome standards, that these standards are open to peer and public review, and that they can be modified if good reasons are put forth.

**Applicability of assessment results**. The second aspect relates to the intended use of the assessment results. For example, if the purpose is to diagnosis student problems, then the assessment must be designed to provide profiles of information that can distinguish among performances for the various concepts or constructs presumed to underlie student difficulties (this is discussed under Validity). For classroom assessment, diagnosis usually implies some form of subtest structure in the assessment, since rarely are assessments designed to assess one unitary concept.

An assessment must provide the necessary discriminations among scores for the scores to be interpreted differentially. This notion is derived from norm-referenced theory, but is appropriate to most classroom assessments, particularly those where relatively fine discriminations among performance levels are made. For example, it is necessary to obtain some differences among individuals if more than one grade is to be given. In theory it is possible to award all students the same grade, but in practice this rarely obtains, and some school systems have policies that prohibit it.

Assessments must be sufficiently precise to provide score differences where different performance levels can be expected, particularly in the region of important cutoff points. Teachers' assessments are rarely designed specifically to give precise measurements at strategic locations on the measurement scale. This is directly related to the problem of score referencing and cutoff points. Since most teacher-made tests cover a number of objectives, students can obtain marks for a variety of skills, some of which are more difficult than, and perhaps even unrelated to, others. Grades and cutoff points are arbitrary at best in this setting, but it is difficult to see how they can be otherwise.

Assessments must be designed to provide the information in a format that makes it usable. For example, if an assessment is to be included with a number of others to obtain a cumulative grade, the assessment must provide results that can be combined in such a way with the results of the other assessments that its relative importance is not subverted. If performance on particular skills is to be evaluated, scores must be obtained for all these skills. Some of these skills may be judged to be more important or of greater significance to future learning than others, leading to more than one standard of performance.

**Communicability of assessment results**. The third aspect refers to the ability of the assessment results to inform the potential users: are the results presented in such a way that users comprehend them readily so that appropriate decisions can be made. This aspect relates to the structure and format of the results, described in the previous paragraphs, but also to the means and the language used to convey them. Consideration must be given to understandings, expectancies, and predilections of the audience(s). For example, the information presented to parents and the public could not be presented with the same kind of technical language as could be that presented to professional educators. As a further example, parents often expect summary percentage or letter grades in reports of their children's progress. These grades have no inherent referential meaning, but alternative reporting formats often require considerably greater explanation and even then may not be accepted by parents.

It is likely that assessments must have different structures depending on their primary purpose, and results on these must be reported in such a way as to provide the major audience with usable information. It is also likely that assessment structure and reporting procedures will vary considerably across grade levels and from one subject area to another.

**Objectivity in scoring**. Objectivity is restricted here to the scoring of students' responses, their products or behaviors. There are many forms of classroom assessment that rely heavily on teachers' judgments in scoring students' work. Many of the important learnings in school must be evaluated by some form of subjective judgement. Subjective assessment leaves room for various types of scorer bias, such as halo effect, leniency/severity error, and logical error. There are a number of procedures to reduce or obviate these forms of error, such as having two scorers judge the same response, but these are not used often in schools (the case studies). Another procedure is to specify clear and detailed scoring guidelines, and to produce a model response or set of behaviors against which to judge students' productions.

Assessment procedures should have some level of objectivity, certainly openness, although this level could vary depending on the context. Some assessment procedures are primarily dependent on professional judgement, procedures such as holistic scoring, for example. It is impossible to avoid subjective scoring of student performances (behaviors or products) since many of the most important learnings cannot be authentically assessed using objective techniques like multiple-choice testing.

**Recommendations for utility of classroom assessments.** There are four recommendations for teacher preparation in assessment that are subsumed under this heading. The first recommendation (13) is the most complex. It refers to score referencing and interpretation. The second (14) and third (15) refer respectively to applying and communicating assessment results. The last recommendation (16) deals with subjectivity in scoring students' work.

**13.** There are no clear, simple guidelines that can be given for the interpretation of scores and other assessment information. It is useful to think in terms of norm-, criterion-, and self-referenced approaches. There are occasions when one approach is superior to another. Prospective teachers must understand the issue of score interpretation in terms of referencing, and be able to apply the relevant techniques in specific situations. In some cases criterion-referencing is the most appropriate, whereas in others it may be necessary to relate scores to the group or some other source of norms. Teachers must also understand that part of the task is setting standards. In some assessments the standards are imbedded in the assessment itself, but there is almost always opportunity for professional judgement. This professional judgement can be enhanced by including the views of others, and by making the process more public.

In part, score interpretations are based on the content of the assessment. But interpretations are also based on the community and school context of the classroom, and the predilections and preferences of the teacher. The school may have policies regarding assessment and grading, which affects how scores are interpreted. Teachers should have a clear sense of how they use assessment information, and how they set standards of performance. They should also be familiar with systematic ways of approaching the problem, how professional judgement is involved and previous performance information can be used, and the normative context in which judgements are made.

**14.** Test scores and their interpretations must be of use for the purposes intended. This means that the appropriate kinds of assessment results and interpretations must be obtainable from the assessment. A simple example of this is subtest scoring (to obtain a profile) versus total test scoring. Teachers must learn how to design the assessment so that the appropriate information is forthcoming. Further, the assessment must produce discrimination of scores at useful points on a scale if the scale is to be converted into grades or mastery-nonmastery categories.

**15.** Communication of assessment results is an important part of evaluation. There are usually a number of audiences, each with different expectations and different abilities to understand the results. It is necessary to simplify interpretations of assessment results in many cases. For example, parents are confounded by too much information if they are confronted by their child's scores on all the objectives in a course, yet this level of detail may be very useful to the teacher and student. It may be necessary also to explain clearly to an audience exactly how assessments have been interpreted and what this means to them. Teachers must be able to communicate assessment results in various ways and to different audiences. These ways would include presenting numerical results and summary statistics, grades, anecdotal or narrative reports, diagnoses, and affective evaluations, and to do this both in writing and orally (e.g., parent-teacher conferences).

Grades and report cards are a systematic way of informing students and parents of student progress. However, the information presented in report cards varies from one school to another. The teacher must be able to provide overall indications of how the student is doing in the class, but also provide supporting information. Teachers must be prepared to meet and work with students and parents, and present and interpret assessment information in this context. Most school systems have some form of parent teacher conference, although this is more prevalent at the elementary levels (Kunder & Porwoll, 1977). Some measurement textbooks provide suggestions and guidelines for conducting these conferences (e.g., Gronlund & Linn, 1990; Nitko, 1983; Sax, 1989).

**16.** Both objective and subjective forms of assessment are necessary in classroom assessment. Teachers can be trained to make some of the subjective assessment procedures more systematic, if not entirely objective. Observations of student behaviors can be designed so that all students are observed under somewhat similar circumstances, for example. The goal is to enhance the quality of subjective assessments, and not to avoid them. Teachers must be trained to use a number of different procedures for subjective evaluations, such as rating schemes, holistic scoring, and narrative descriptions. These procedures can be grounded in subject areas, where particular approaches may be prominent (e.g., holistic scoring in language arts; Huot, 1990).

Teachers should be aware of potential biases that may occur in subjective evaluation, and particular situations where they may be more prominent (e.g., halo effect and logical error in observing students in group activities). They should also be aware of personal characteristics and beliefs that they themselves might have that could adversely affect scores. There are a number of procedures that can be invoked which reduce the potential for scorer bias. These are described in many measurement textbooks.

## Efficiency

Efficiency covers questions such as, Are the evaluations efficient in terms of teacher and student time and effort? In terms of teacher and school time for preparation of the assessment? In terms of classroom time for administering the assessment? Efficiency is important to classroom assessment, particularly since teaching time is at a premium and a considerable amount of this time is taken up with assessment. The notion of efficiency refers to the use of teacher and student time to obtain the assessments necessary for particular purposes (it does **not** refer to efficiency in the statistical sense.)

The characteristic of efficiency is restricted to those aspects of the assessment process that are concerned with obtaining information at reasonable "cost". Some of the criteria identified by Tuckman (1975) and Gronlund (1985) fall into this category, but others are more appropriately designated as utility. Three main aspects of the assessment process have been placed under this heading: ease of preparing the assessment (e.g., test development), convenience of administration (ease and practicality of administration, time required for administration), and ease of obtaining scores and reports (e.g., time required for scoring, recording, interpreting, and reporting). The notion of efficiency is relative in two senses: an assessment is efficient relative to its purpose and the intended audience, and it is also relative to other, alternative assessments that potentially could yield the necessary information.

**Recommendations for the efficiency of classroom assessments.** There are more and less efficient approaches to assessment. Some are exceedingly demanding of teachers' time, such as written analyses of student essays. The purpose of the assessment and the demands of the situation dictate how much effort can be expended on the assessment.

**17.** Since both teacher preparation time and classroom instruction time are at a premium, it is incumbent on teachers to be as efficient as possible in conducting their assessments. The amount of assessment effort on the part of the teacher and of the students should be roughly proportional to the significance of the evaluation and the consequences of the decision. Teachers must know the approximate effort and time required for various assessment procedures. They must also know ways to speed up the process. There are three general categories of assessment procedures that should be understood in this regard, applied performance assessment including observation, selection-type tests such as multiple choice, and longer constructed response assessments such as written papers and research reports.

Some assessment approaches can be conducted with little disruption to classroom instruction. Examples include teacher marking of student work, in-class observations of student skills and behaviors, and peer evaluation of student presentations. Some of these require considerable before-class preparation time, such as applied performance assessment, whereas others require after-class marking on the part of the teacher (e.g., written assignments). The out of class effort must be balanced against the benefits of these types of assessment. There is usually a trade-off that must be made with any approach to assessment. For example, multiple-choice items can take considerable time to produce but they can be marked rapidly, whereas the opposite is true of essay questions.

## Summary of Recommendations for Teacher Preparation in Classroom Assessment

In the previous section of this chapter, the characteristics of good classroom assessment practices were discussed under four major headings, Reliability, Validity, Utility, and Efficiency. Within these broad categories 17 recommendations for teacher preparation in classroom assessment were identified. The characteristics were summarized, each recommendation was given a heading and summarized, and these were then collated in a document. This summary and the recommendations form the focus of the model for teacher preparation in classroom assessment. The document was presented to educators for their review; it is duplicated in Appendix D. The recommendations are also summarized below.

### Reliability--Recommendations 1 to 5

Reliability refers to the consistency of assessment results, to the stability of scores obtained from using assessment procedures. This implies that the overall procedures for an assessment should be included in considering the reliability. There are five recommendations regarding reliability for preparation of teachers in classroom assessment.

### 1. The importance of reliability for high-stakes assessments.

Reliability should be greatest for those assessment results that have the greatest consequences (i.e., high-stakes assessments). If the purposes of the assessment have considerable consequences for students (or even to what happens in the classroom or school), then care must be taken to ensure reliability. To understand the importance of reliability for high stakes assessments, teachers must be aware of the implications that low reliability might have for various decisions. The most significant consequences for classroom assessments are usually related to grading and reporting of student progress, with promotion-retention decisions having the most dramatic effects on students. It often takes considerable effort to enhance reliability in classroom assessment, so teachers must be able to make trade-offs and enhance reliability where it is most imperative.

### 2. Practical ways of improving reliability in classroom assessments.

Reliability can be enhanced in two general ways which are practical in the classroom: making the assessment procedures more explicit and systematic, and increasing the amount of high-quality information gathered. Teachers must know how to enhance the reliability of important types of assessment: subjective observations, constructed response testing, objective testing, portfolio assessment. Each has its unique problems, but the reliability of each can be enhanced. The first principle implies that assessment procedures should be clearly understood by students, and that it should be applied systematically to all students and to all occasions. The second principle requires that collecting more assessment information may improve reliability of the overall information base, but the additional assessment information must have some inherent reliability. Teachers should be aware of factors that contribute to low reliability (e.g., Frisbie, 1988; Traub & Rowley, 1991).

### 3. Increasing reliability through multiple quality assessments.

Students must have the opportunity to exhibit their learning in several ways and under differing circumstances, particularly if the results for a student are not unequivocal. More assessment alone is not the answer, but judicious choices of assessment procedures should be made. This recommendation extends on 2 above, but emphasizes that teachers must be able to obtain assessment information in differing ways and settings to ensure that the information on students is accurate.

### 4. Numerical procedures for estimating reliability in classroom assessments.

There are a number of numerical procedures that are simple enough to calculate so they can be used to help determine if classroom measurements are reliable. With modern computing equipment being readily available, it is becoming reasonable for teachers to check the reliability of their assessment information periodically, particularly in high-stakes situations. There are several fairly simple statistics which are applicable a variety of scoring systems, such as $p_0$ which is the proportion of consistent categorizations of students using two parallel assessment procedures. There are also programs available on microcomputers to compute a variety of test statistics, such as means, standard deviations, and internal consistencies (e.g., K-R20 formula). It may be necessary to use internal consistency statistics, such as split-half reliability procedures, since there is usually only one assessment. Teachers must know how to use some of these methods and interpret the results (e.g., Frisbie, 1988; Gronlund & Linn, 1990; Nitko, 1983).

### 5. Reliability of subjective forms of assessment.

Subjective forms of assessment are an essential part of teaching. This includes assessments of student products as well as observations of behaviors. Therefore, teachers must be able to develop and use systematic marking procedures (e.g., for essays, projects). As well, they must prepare and use effective observation techniques. Teachers must understand that subjective forms of assessment are prone to unreliability and to personal bias. They must also know practical procedures which can ameliorate these problems, and enhance the reliability of the assessment information. This includes specifying observation and rating schemes, designing systematic schedules, recording observations, and conducting consistency checks on the procedures.

## Validity--Recommendations 6 to 12

Validity is the most important characteristic of assessment. It refers to the "correctness" of the results of an assessment. The unified view of validity, which is generally accepted today, implies that both logical and empirical evidence must be amassed, and that the consequences of the assessment must also be considered (e.g., Cronbach, 1988; Messick, 1989a, 1989b; American Educational Research Association et al., 1985). The validation process has substantive, structural, and external components, and all of these are applicable to classroom assessment. No longer is content or curricular validation considered sufficient.

There are seven recommendations relating to validation of classroom assessment and teacher preparation. These encompass the three components of validity, but highlight aspects that are of particular importance in the classroom, and yet are practical.

### 6. Increasing assessment of important, higher level cognitive learnings.

Classroom assessment procedures must provide information on important student learnings which include more complex and higher-level thinking. Therefore, teachers must be able to design and conduct classroom assessments of both the content and thinking processes identified in subject areas and the broader thinking skills, communications skills, and cognitive strategies (metacognition) that are presently emphasized by our schools. Important higher level learnings have been clearly identified through curriculum guides, teaching materials, and professional documents and journals. Preparation of teachers would include general skills in developing assessments to tap higher-level thinking abilities and affective learnings (e.g., Stiggins, Rubel, & Quellmalz, 1988; Wolf, Bixby, Glenn, & Gardner, 1991), as well as skills to produce assessments particular to various subject disciplines (e.g., in science, Yager, 1989). As an example, general skills would include such things as preparing clear instructions for tasks given to students and developing scoring procedures for written work, whereas the language arts discipline would give rise to the skills in applying holistic, analytic, and descriptive assessment of prose passages.

### 7. Assessment must be designed to support and inform good instruction.

Classroom assessment procedures must reflect and support good pedagogy, by emphasizing what is important and setting standards for its attainment and to provide practice for this kind of thinking (e.g., Wiggins, 1989a, 1989b). Teachers must have training and actual experience in preparing and using assessment procedures that support the most important learning in classrooms, those that should be emphasized in class and on the assessments. Teachers must know how to develop and use a wide variety of assessment procedures, such as paper-and-pencil techniques, but more importantly, they should include systematic observation and applied performance assessment. This recommendation extends upon the previous one. Both the focus of the assessment and the style of the assessment must approach that which formed the learning. Further, teachers must be able to focus the assessment on the intended learnings, and not to include extraneous factors.

### 8. Assessment must support various purposes, including instructional diagnostics.

Teachers must have the training to devise assessment procedures for a variety of purposes, such as for grading of students, individual and group instructional diagnosis, and evaluation of instruction. There are other purposes, but these are likely to be secondary. Teachers must also know when and how it is possible to accommodate more than one purpose, and to recognize when this leads to incompatible procedures. Diagnosis requires that there be detailed information on meaningful diagnostic categories (skills, abilities), whereas summative assessments may sample from a number of content and skill domains. Although it is not practical to prepare teachers for individual diagnosis, or even instructional diagnosis, in all subject areas, teachers should understand the general principals and procedures, and be able to apply them in selected settings.

### 9. Empirical validity checks of the skills covered in classroom assessments.

Several tasks or items should be devised to assess the same skill or concept and these should yield similar results with students. A check for this presumed convergence is one way of testing if there is a stable, meaningful underlying construct, and that scores can be interpreted. Teachers must know how to check for this type of homogeneity, be it for total assessment scores or for subtest scores. This involves identifying the assessment tasks or items which should "hang together" and checking student results to see if this is the case. This recommendation extends upon the previous one by specifying that teachers know procedures by which they can empirically check whether the assessment does provide diagnostic or other information, and that scores can be interpreted with some confidence for the purposes intended. For example, homogeneous items should correlate with one another, and this can be checked by looking at the performance of each item in relation to the other items flagged for the same skill or objective.

### 10. Making assessment results reflect the focus and purpose of the assessment.

Classroom assessments must provide the scores and information so that they reflect on the features and content of importance, and thereby assist in making the decisions for which they are intended. Teachers must be able to obtain scores from assessments corresponding to the nature and purpose of the assessment. This involves identifying clearly in advance of producing the assessment the approximate importance, and hence the weighting, of various areas of content and levels of skills (essentially producing a table of specifications for the assessment). Recommendation 10 relates to recommendations 7 and 8, but emphasizes more specifically the procedures by which scores are obtained, either from direct ratings, judgments based on observations or products, or from composites of assessment items. The scores must reflect clearly the main intent of the assessment, and should not be subverted by such procedures as inappropriate weighting of items or assessment components (Terwilliger, 1989; Thayer, 1991). This is particularly important in summative evaluation where composite scores are obtained from a variety of assessment sources. Procedures that are useful in this regard are the careful specification and weighting of aspects to be included in the assessments (e.g., test blueprints), and the explicit formulation of how composites are to be obtained.

### 11. Confirming the results of assessments by comparison with other assessments.

Results of previous assessments or other relevant evaluative information provide a check on assessment results, thereby giving external evidence for their validity. Teachers must base their judgements of students on information from multiple, diverse sources. The validity of the information should also be checked periodically in this way. This can be done informally by comparing the results of particular students on one assessment to those on another, or by using systematic techniques, such as correlation. It is unlikely that statistical procedures will be used in the classroom setting, but teachers should understand the importance of external validation. They should attempt to obtain several assessments of students' learning using different assessment procedures, and compare students' performances on these to see if there are glaring discrepancies.

### 12. Removing bias and prejudice in classroom assessments.

Classroom assessments should be free of biases and prejudices that may have negative impact on students, insofar as these can be identified from our understanding of subgroup differences. One issue is that of fairness or equity: students should have equal opportunity to achieve various results, dependent only on the skill or content which is the focus of the assessment. A second issue is more subtle: it is bias or prejudice that has a negative reflection on particular subgroups and may over the long run have a negative impact on student learning and on these subgroups. Teachers should be aware of the common forms of bias that are reflected in stereotypes of subgroups based on characteristics such as gender, ethnicity and race, socioeconomic status, and geographical location, and they should be experienced in use of guidelines to reduce this bias (e.g., Shepard, 1982; Tittle, 1982). Teachers should submit their assessments to peer and public review. More subtle forms of bias exist in such things as language structure and complexity, and forms of response required of students. To reduce potential bias of this nature the performance of identified subgroups can be compared across various techniques of assessment.

## Utility--Recommendations 13 to 16

The ability of assessments to provide information that is readily and appropriately interpreted is described under the topic of Utility. This includes such aspects as the referential basis for the assessment scores, the clarity of the communication, and the objectivity of the scoring process. There are four recommendations under this heading.

### 13. Using score referencing to make value interpretations of assessment results.

The interpretation of assessment information is a difficult and value-laden problem. There are no clear, simple guidelines that can be given but there are three general approaches to interpreting scores: norm, criterion, and self referencing. Prospective teachers must be able to apply the relevant techniques in specific situations and for certain purposes. Teachers must also understand that part of the task is setting standards. This almost always involves professional judgement, which can be enhanced by including the views of others and by making the process more public. The standards of performance or the norms of behaviour are determined by such factors as the subject matter and the purpose of the educational program, the policies and context of the school and school system, and the personal and professional views of the individual teacher. Teachers must be able to explicate their standards and how they make value interpretations.

### 14. Obtaining the scores necessary to use the assessment results properly.

For test scores to be useful, the appropriate kinds of assessment results and interpretations must be obtainable from the assessment. This relates to Recommendations 8 and 10 under the topic of validity, and extends upon the notions of producing scores and score profiles. Further, the assessment must produce discrimination of scores at useful points on a scale if the scale is to be converted into grades or mastery-nonmastery categories. The problem of dealing with scores that are close to important cutoff points was discussed under Recommendation 3. However, teachers should be aware of the potential problem in advance and devise assessments which have the potential of discriminating at important points on the scale (e.g., at scores near 80 when 80 and above is awarded an "A"). One way to do this with paper-and-pencil testing is to design items with the appropriate range of difficulty levels.

### 15. Communicating the results and interpretations of assessments.

Communication of assessment results is an important part of evaluation in the classroom. It is usually necessary to simplify interpretations of assessment results for reporting to individuals outside the school system. For example, parents are confounded by too much information if they are confronted by their child's scores on all the objectives in a course. However, this may be very useful information to the teacher and student. It may also be necessary to explain clearly to an audience exactly how assessments have been interpreted and what this means to them. Teachers must be able to communicate assessment results in variety of ways and to different audiences. The most common form of formal reporting in schools are student report cards. There is no one best way of reporting student learning, but there are suggestions of procedures that are more effective than others. Teachers must also know how to communicate directly with parents on their children's progress, such as in parent-teacher interviews.

### 16. Controlling the effects of subjectivity in observing and marking.

Subjective assessment procedures can be made more systematic. Observations of student behaviors can be designed so that all students are observed under somewhat similar circumstances, for example. Teachers must be trained to use a number of different procedures for subjective evaluations, such as rating schemes, holistic scoring, and narrative descriptions. The problems of reliability in subjective scoring are discussed under Recommendation 5. The problem highlighted here is that of effectively transferring subjective judgements into statements that can be understood clearly. Often these statements are in the form of numbers, such as in holistic scoring of writing. But there are other ways that qualitative judgements can be communicated with clarity. Teachers must be familiar with ways of conducting subjective assessments and reporting them effectively.

## Efficiency--Recommendation 17

The last recommendation deals specifically with the relative cost in time and effort of assessments of various kinds and for different purposes (Recommendation 17).

### 17. Making maximum use of assessment time and effort.

Teacher preparation time and classroom instruction time are at a premium; therefor, it is necessary for assessment to be conducted as efficiently as possible. The amount of assessment effort on the part of the teacher and of the students should be roughly proportional to the significance of the evaluation and the consequences of the decision. Teachers must know the approximate effort and time required for various assessment procedures. They must also know ways to speed up the process. There are three general categories of assessment procedures that should be understood in this regard, applied performance assessment including observation, selection-type tests such as multiple choice, and longer constructed response assessments such as written papers and research reports. There are ways that the efficiency of each can be enhanced.

# V. EDUCATOR REVIEW OF THE RECOMMENDATIONS FOR TEACHER PREPARATION IN CLASSROOM ASSESSMENT

In the previous chapter, 17 recommendations for teacher preparation in classroom assessment were identified from the classroom assessment literature and the case studies. The recommendations were categorized according to four characteristics considered central to assessment: reliability, validity, utility, and efficiency. These served to organize the recommendations on the basis of the measurement and evaluation theory. The context for classroom assessment and the realities of teacher training, as well as current knowledge on classroom practices and teacher skills, were also considered in formulating the recommendations. The purpose of the study was to provide a model for teacher preparation in classroom assessment that was effective yet practicable given the context of teacher education. Therefore, advice was sought from a number of educators to determine which recommendations should be enacted, and how this should be done.

A procedure was developed for a panel of educators to review the recommendations for their applicability to preparation of teachers and the teaching context, and their appropriateness to measurement, curriculum, and pedagogical principles. Further, this review was to suggest practical approaches to preparing teachers and how this aspect of teacher preparation was to correspond with other aspects of teacher preparation. A number of educators representing different functions in school systems, and with varying backgrounds and interests, were invited to participate in the review.

The chapter begins with a discussion of the basis for the review. The model was to be oriented towards initial preparation of teachers for classroom assessment, but this is discussed in the broader context of teacher professional preparation and ongoing development. This is followed by a description of the procedures used to focus the responses of the reviewers. The results of the review are summarized next, and from this suggestions are made for implementation of the recommendations. The chapter concludes with a section on implications for instruction in classroom assessment: the model.

## Basis for Reviewing the Recommendations for Teacher Preparation in Classroom Assessment

Evidence from observations of how teachers assess in the classroom (Chambers, 1982; Haertel, 1986; Stiggins, 1986a, 1988b; Stiggins, Griswold, & Wikelund, 1989; the case studies) and from interviews and surveys of teachers (Gullickson, 1982, 1985; Stiggins & Bridgeford, 1985; Webster, 1987) clearly implies that teachers typically do not apply many of the principles of good assessment procedures in their classrooms. Other evidence suggests that teachers do not understand many of these principles well, although some of the teachers in these studies had previous training in tests and measurement (e.g., Boothroyd, McMorris, & Pruzek, 1992; Carter, 1984; Chambers, 1982; Marso & Pigge, 1992; Newman & Stallings, 1982). There appears to be need to revise and extend initial teacher preparation in classroom assessment (e.g., American Federation of Teachers et al., 1990; Barnes, 1985; Gullickson, 1986b; Nitko, 1991a; Rogers, 1991; Schafer, 1991; Schafer & Lissitz, 1987; Stiggins, 1988a, 1991a), and also to provide practicing teachers with relevant inservice work and other forms of continuing education (e.g., O'Sullivan & Chalnick, 1991; Stiggins, 1991a).

Teacher development can be categorized as initial education, termed preservice or precertification, and continuing education or inservice professional development (Lanier

& Little, 1986). Teachers can become better equipped for classroom assessment through preservice or inservice education, or a combination of the two. Some of the knowledge and skills teachers need to conduct good assessments can be taught as part of initial teacher education, but probably some must be developed during the time a teacher is involved in actual teaching.

Those responsible for teacher education programs, teacher educators, must determine whether the focus is to prepare teachers in broad, theoretical bases of major disciplines of knowledge and their applications in education, or to provide them with the procedural skills and techniques of the teaching craft. Lanier and Little (1986) describe this as distinguishing the approach to teacher education as liberal-professional or technical-professional. Both approaches acknowledge the importance of perceiving teacher education as professional, which includes preparation related to the ethos and culture of the profession (Lortie, 1975), rather than purely academic, which emphasizes the subject discipline approach as commonly exemplified in faculties of arts and sciences. The liberal-professional approach emphasizes the "intellectually deep and rigorous study" of education from various vantage points having their methodological and substantive roots in disciplines such as the arts and sciences, including history, philosophy, and sociology. The technical-professional approach favours training of teachers in prescriptive knowledge and skill performance, these being based on process-product research evidence (outside-expert oriented) and on "tried-and-true" classroom techniques (authority oriented, and conformist). Lanier and Little (1986) argue that at present teacher education programs tend to the technical-professional approach, and conclude that:

> The increasing proportion of career teachers makes the oft-repeated call for a liberal-professional approach to teacher education all the more persuasive. . . . preparing career teachers for their continuing education requires greater emphasis on liberal-professional studies than is presently the case. . . . Unfortunately, changes in the teacher education curriculum have tended to move it in the opposite direction, giving increased dominance to the mastery of skills with immediate practical value. What is worse, studies of the curriculum of initial and continuing teacher education show it to be fragmented and shallow. (p. 555)

There is no one best approach to teacher education. However, there are issues that can be addressed that lead to a reasoned adoption of an approach, issues of liberal versus technical education of teachers, of general knowledge versus specific skills, and of integrating the liberal aspects of teacher education with those of a professional character. This has implications for any program that purports to develop teachers' classroom assessment skills. For example, should teachers be taught the underlying theory and principles of assessment, with their attendant complexities, or should teachers be taught a narrow band of directly applicable assessment techniques?

It may not be practical, let alone reasonable, to expect prospective teachers to spend a large amount of their university time in developing an in-depth understanding of the broad field of measurement and evaluation, much of which would not be applicable to their future teaching situations. The amount of preservice time devoted to direct study in education is usually only a fraction of the overall initial education of teachers. Initial education generally consists of four or five years of formal coursework in general-liberal studies, in major and minor teaching areas, and in pedagogical study. It is estimated in the United States that "the course work in pedagogical studies generally represents only about one-fifth of a secondary teacher's required program and about one-third of an elementary teacher's program" (Lanier & Little, 1986, p. 529). Requirements differ considerably from one university to another and among programs within a given university, such as between a four-year BEd program and certification after a first degree.

As an example, at the University of Alberta for 1990/91, of the 120 course weight requirements for a BEd degree 48 (40%) must be chosen from outside the Faculty of Education for the elementary route and approximately 54 (45%) from outside for the secondary route. However, some of the course requirements within the Faculty of Education would be considered foundational and not directly pedagogical (termed basic education); courses such as those in Educational Foundations and Educational Psychology, which may account for as many as 30 (25%) course weights of an elementary BEd program. Furthermore, a student may be certifiable if she/he obtains a degree from another faculty and completes 30 (25%) course weight requirements in Education. Initial teacher education in classroom assessment must be considered relative to the total amount of time in the teacher preparation program and to the amount of time that can be devoted to educational studies.

Although there are many aspects of classroom assessment which are appropriate to all levels and specialty areas for which teachers are prepared, it is unlikely that there would be much motivation for all students of education to be extensively trained in a common set of specific assessment skills. For example, one technique of considerable direct utility to teachers of language arts is holistic and analytic scoring of a variety of forms of student writing. However, students interested in teaching the sciences at higher grade levels may not be expected to use these skills to any great extent, and what may be of greater relevance to them would be learning how to effectively assess science process and laboratory skills. What is more likely to be useful for all teachers is to be familiar with, but not extensively trained in, a number of procedures for the scoring of written material, including the application of selected procedures and their particular strengths and weaknesses. It would be desirable, however, for students in particular teacher training areas, such as language arts, to receive more extensive experience in this method of evaluation. It could be expected, then, that some aspects of assessment should be part of initial education for all prospective teachers whereas other components should be made specific to the anticipated teaching areas of particular groups of education students.

Present thinking suggests that teacher socialization as professionals, is much more part of teacher experience in schools than of initial education (Feiman-Nemser & Floden, 1986). These authors concluded that there is no one generalizable school culture, and variations exist among schools and among subgroups within a school. This implies that it would be difficult, and perhaps futile, to impart much of the understanding of the culture of teaching in initial teacher education. It also implies that some of the modes of teaching and norms of practice are developed while teachers are in the classroom. This suggests that at least some of the detailed practices pertinent to teaching in particular contexts would be more a matter of continuing education than of preservice education. Initial teacher education may only reasonably hope to achieve the more "liberal" goals of education.

A number of solutions have been proposed for enhancing teacher skills, and this has been for preservice preparation of teachers (e.g., Gullickson, 1986a, 1986b; Gullickson & Hopkins, 1987; Nitko, 1991a) and for inservice professional development (e.g., Stiggins, 1991a). However, the emphasis here was to place assessment in a general model for the initial education of teachers, although reference could have been made to possible continuing education where this appeared appropriate, and the questions posed for the review reflected this. Initial teacher preparation in assessment could be the focus of a course, which is common in many university programs. Or it could be brought about in other ways: for example, assessment could be integrated with instruction in curriculum and pedagogy. The task of the review panel was to determine the focus of this instruction and what approach or approaches would be best.

## Procedures to Review the Recommendations

Reviewers were asked to rate the importance of each of the 17 recommendations. As well, four features related to the delivery of instruction were identified to focus the reviewers' responses. A document was prepared that explained the basis for the recommendations and the intent of the review, a summary of the characteristics and the recommendations, detailed procedures for the reviewers to use, and forms for the reviewers to respond to each recommendation (see Appendix D). This document was presented to the reviewers, who were asked to respond according to the procedures identified, but also to give their opinions and views more generally.

## Individuals Selected to Review the Recommendations

The model was intended to present an approach to the development of prospective and practicing teachers that is realistic and practicable. Therefore, to review the recommendations, individuals were selected who were deemed knowledgeable in various aspects of the education process, including teacher preparation and classroom assessment. Reviewers were identified from each of four groups.

1. **Educational measurement and evaluation specialists (4).** These persons were to assure that the recommendations for preparing teachers conformed to good assessment practices. Three university professors were identified with specialized training in educational measurement, and with experience in teaching measurement to prospective teachers and in professional development of practicing teachers. Also, one measurement consultant (at the board level) was identified who had specialized training in educational measurement, and who worked with teachers in the preparation of district-wide tests and in their professional development in assessment.

2. **Curriculum specialists (5).** These persons were to determine that the measurement focus outlined in the recommendations conformed to what was thought to be good assessment practice in various curricular areas, and to be good pedagogy and instructional practice. There are many subjects that are taught in schools, but one guiding feature was to include persons with experience in subject areas that are taught at most grade levels and that represent somewhat different disciplines of study: that is, humanities and sciences. The case studies focused on teachers of science and social studies at the junior high level. Three university curriculum educators were identified, two with specialization in social studies and one in science. All had experience as teacher educators in the area of specialization and in the professional development of practicing teachers. Two curriculum specialists in science, who worked as consultants at the board level, were identified; they were involved directly in teacher professional development.

3. **School- and division-level administrators (2).** These educators provided experience with the supervision of teachers that would help assure that the assessment procedures would be administratively feasible, appropriate in the classroom, and reflective of effective practices with teachers. These persons were familiar with what teachers do in the classroom, and with the needs and concerns that teachers have. This included one school vice-principal with an interest in classroom assessment, and who had considerable experience in teaching measurement procedures to teachers both at the preservice and inservice levels. It also included one school division/district administrator (assistant superintendent) who had responsibilities in teacher supervision and professional development, and who had considerable experience in measurement and evaluation.

4. **Practicing, experienced teachers (5).** These persons were responsible for ensuring that the assessment procedures would reflect the realities of the classroom

and be practical for teachers. The teachers were experienced with the subject matter at a number of grade levels. Four of these teachers were the ones involved in the case studies. Two of these taught science at the junior high school level as their primary responsibility, and two taught social studies. An additional teacher of social studies was included, since one of the two social studies teachers also taught mathematics and may not have been as reflective of teachers of social studies generally.

**Others with experience in classroom assessment.** Other educators could have provided useful input into teacher training for classroom assessment, but those noted above were deemed adequate to provide critical feedback to the recommendations.

## Features for Review of the Recommendations

Approaches to instruction vary in many ways, one of which is the identification and selection of the goals and objectives of the instruction. There are a number of important characteristics of classroom assessment that all teachers, irrespective of their specialties, may be expected to understand, and probably much of this could be taught to prospective teachers prior to their having actual classroom teaching experience. This would include certain skills, and some issues and concerns with which all teachers should be familiar. Some of this material may be suitable to teach to groups of prospective teachers who are heterogeneous in their backgrounds and interests. Other aspects may not be as applicable to students overall, and may require a differentiated approach in teacher education. There are also many different ways to teach the assessment material. Four features of teacher education were identified that relate to the focus, organization, and delivery of initial instruction in classroom assessment. These features directed the review of the 17 recommendations proposed for the model. They were not considered to be discrete, and suggestions given for one feature could have implications for others.

For the review it was assumed that prospective teachers (students in education programs) would have adequate substantive knowledge and understanding in their specialty, if they are subject specialists. It was assumed further that they would have the knowledge and appropriate background for the grade levels of interest, and an understanding of principles related to curriculum, child development and learning, and classroom organization and management. Thus, the review focused on the nature of preparation in assessment, and not on the expected outcomes of schooling in content areas and at various grade levels.

**Structured procedures for the review.** The characteristics of good classroom assessment and the 17 recommendations (presented in Chapter IV), as well as the features that formed the focus for the review, were described in the document that was presented to the educators who were to review the recommendations (see Appendix D). The reviewers were asked first to rate the importance of each recommendation for initial teacher education, and then to review it in terms of four features of instructional delivery of classroom assessment skills and knowledge related to that recommendation. The instructions given to reviewers are repeated below, and the structured scales and response forms are presented in Figure 2. Also, reviewers were asked to comment on each recommendation with respect to the features and more generally.

The actual instructions to reviewers for rating the importance of each recommendation were:

> **Importance:** The recommendations are not of equal significance, nor is it expected that they would be given similar amounts of effort in a teacher education program. Your view regarding the importance of the recommendation should be

noted, and you may wish to comment on the relative amount of emphasis that should be given to it in teacher preparation.

Check whether you think it is "Very important" or "Not at all important" or somewhere in between.

The first feature was intended to identify the amount of substantive and theoretical development necessary for instruction in assessment, and the level of specific technical prescription that would be desirable. This distinguishes between the substantive background of evaluation theory and methodology (e.g., statistical underpinnings of measurement) and the actual assessment practices suggested for classroom teachers (e.g., context dependent items, holistic scoring of writing). Instructions to reviewers were:

> **1. Level of Specificity:** Amount of substantive and theoretical development, and level of specific, technical prescriptiveness of the instruction in a teacher education program.
>
> Check whether you think it should be "Theoretically based and developed from general principles" or "Specific and providing classroom oriented prescriptions", or somewhere in between.

**Figure 2.** Form for Reviewers' Structured Responses to Each Recommendation

| | | | | | |
|---|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important | |
| (check) | *(4)* | *(3)* | *(2)* | *(1)* | |
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | *(4)* | *(3)* | *(2)* | *(1)* | ____ |
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | *(4)* | *(3)* | *(2)* | *(1)* | ____ |
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ____ | ____ | ____ | ____ | ____ |
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ____ | ____ | ____ | ____ | ____ |

Note. Where numerical scales appeared appropriate the number is given in parentheses above the scale position; the numbers were applied after reviewers had responded.

The second feature related to the type and level of differentiation of the material and its delivery. Differentiation could be based on the specialization of the prospective teachers; the grade and age level emphasis (e.g., early, middle, senior years; adult learners); the subject and discipline specialty (e.g., natural sciences, mathematics and related fields, social sciences, humanities, languages and language arts, fine arts); and the program specialization in education (e.g., special education, gifted learners, mentally handicapped, counseling and support services). The instructions to reviewers were:

> **2. Common or Differentiated:** Type and level of differentiation of the material and its delivery in a teacher education program, considering such characteristics as specialization.
>
> Check whether you think it should be "Common for all students (prospective teachers)" or "Differentiated for particular groups", or somewhere in between.

The third feature focused on the method and timing of delivery of the instruction, and its relationship to the overall teacher education program: for example, should instruction be in a separate course, should it be in short or mini courses related to specific topics but placed within the broader program and coursework of teacher education, or should the material be integrated into other teacher education courses; when should the material be taught in the teacher education program; and how should the material be articulated with other aspects of the teacher and specialist education programs (and also programs with faculties outside education). The instructions given to reviewers were:

> **3. Method of Delivery:** Method and timing of delivery of the instruction, and relationship to the overall teacher education program.
>
> Check whether you think it should be "Part of one course in measurement", "Short course", "Seminar", or "Part of a course in curriculum or pedagogy", or some combination of these.

The last feature identified for reviewers was the nature of the learning experiences for students. Reviewers were presented with:

> **4. Nature of Instruction:** Nature of the learning experiences for students in a teacher education program.
>
> Check whether you think it should be "Lecture", "Laboratory", "Clinical practice", or "Practical, in-school experience", or some combination of these.

Several potential constraints were noted in the document for reviewers, such as the limited time available in a teacher education program and the course structure of the programs at most universities. Possible content topics for a teacher education program in classroom assessment were outlined so that reviewers would have some idea of how content might apply to the more broadly stated recommendations. These topics reflected the seven assessment competence standards identified for teachers in the United States (American Federation of Teachers et al., 1990) but did not parallel them. The topics were not exhaustive of all possible content, nor a course outline, but gave an indication of what might be addressed in preservice instruction in classroom assessment for teachers.

Analysis of the responses to the structured procedures. All of the 16 educators invited to review the recommendations responded to the document. Of these, 15 used the structured format provided, and their responses were tabulated and summarized. The sixteenth reviewer commented on the recommendations but did not respond according to the format outlined in the review document.

For the importance rating, and for Features 1 and 2, points were designated later for the various positions on the scale (indicated in parentheses on the scales in Figure 2). These points were used for computing means and comparing these for the 17 recommendations. Analyses were conducted to determine for the 17 recommendations which of them differed in mean rating of importance, and whether their mean responses to Features 1 and 2 differed from a theoretical middle point of 2.5 (midway between the scale extremes of 4 and 1). For Features 3 and 4 there were no meaningful underlying scales, and response positions were treated categorically. Analyses were conducted to determine for each of the 17 recommendations whether the proportions of responses to the four categories for each feature differed from what might be expected by chance.

The analyses of the importance ratings and of the responses to the four features are summarized below. Details are given in Appendix E. The recommendations were ordered and grouped on the basis of the importance ratings, and this was used to organize the discussion that follows on the implementation of the recommendations. Included are reviewer comments. These discussions incorporated the characteristic of assessment that each recommendation addresses: reliability, validity, utility, and efficiency.

## Review of the Recommendations

Reviewer ratings of the importance of the 17 recommendations are discussed first, followed by the results on the four features for instruction. In the document presented to the reviewers a summary statement was prepared for each recommendation. To facilitate the discussion these are given below in Figure 3. The summary statement served only to reflect what was generally intended by a recommendation, and the recommendations to which reviewers responded were expressed in greater detail and clarified (Appendix D).

The importance ratings given to recommendations were compared using a two-tailed *t* test for correlated samples based on the average variance and covariance. These tests were problematic since there were 136 comparisons and the tests were not independent. However, the number of respondents was too small to apply multivariate statistics. But it was possible to identify what difference could be considered important in a statistical sense if a pair of means was taken in isolation. Two-tailed *t* tests were also used to determine if the mean of responses to Features 1 and 2 for each recommendation differed from 2.5. These tests, too, were not independent for the 17 recommendations, but provided some guidance for interpretation of the results.

There was no meaningful scale for Features 3 and 4 so frequencies were compared across the response categories for each recommendation using a $\chi^2$ goodness-of-fit test based on the null hypothesis of equal frequencies of responses for each category (25% of responses). As with the tests for the other features, the $\chi^2$ tests were not independent for the 17 recommendations. Also, some of the responses to one recommendation clearly were not independent since often one reviewer selected more than one category. Further, expected frequencies approached 5, which is the minimum suggested for goodness-of-fit tests (e.g., Ferguson & Takane, 1989; Shavelson, 1988). Nevertheless, these analyses were used to help guard against over interpretation of perceived differences.

Since the number of respondents was small, all of the statistical analyses must be viewed cautiously, but they gave a basis for deciding which recommendations were considered more important, and which instructional approaches were thought to be most appropriate for a recommendation. Appendix E provides a further discussion of these analyses as well as details of the results.

**Figure 3**. Summary Statements of the 17 Recommendations in the Order They Were Presented to Reviewers

**Reliability**

1. The importance of reliability for high-stakes assessments.
2. Practical ways of improving reliability in the classroom.
3. Increasing reliability through multiple quality assessments.
4. Numerical procedures for estimating reliability in classroom assessments.
5. Reliability of subjective forms of assessment.

**Validity**

6. Increasing assessment of important, higher level cognitive learnings.
7. Assessment that supports and informs good instruction.
8. Assessment for various purposes including instructional diagnostics.
9. Empirical checks of validity of the skills covered in classroom assessments.
10. Making assessment results reflect the focus and purpose of the assessment.
11. Confirming the results of assessments by comparison with other assessments.
12. Removing bias and prejudice in classroom assessments.

**Utility**

13. Making value interpretations of assessment results using referencing procedures.
14. Obtaining the scores necessary to use the assessment results properly.
15. Communicating the results and interpretations of assessments.
16. Controlling the effects of subjectivity in observing and marking.

**Efficiency**

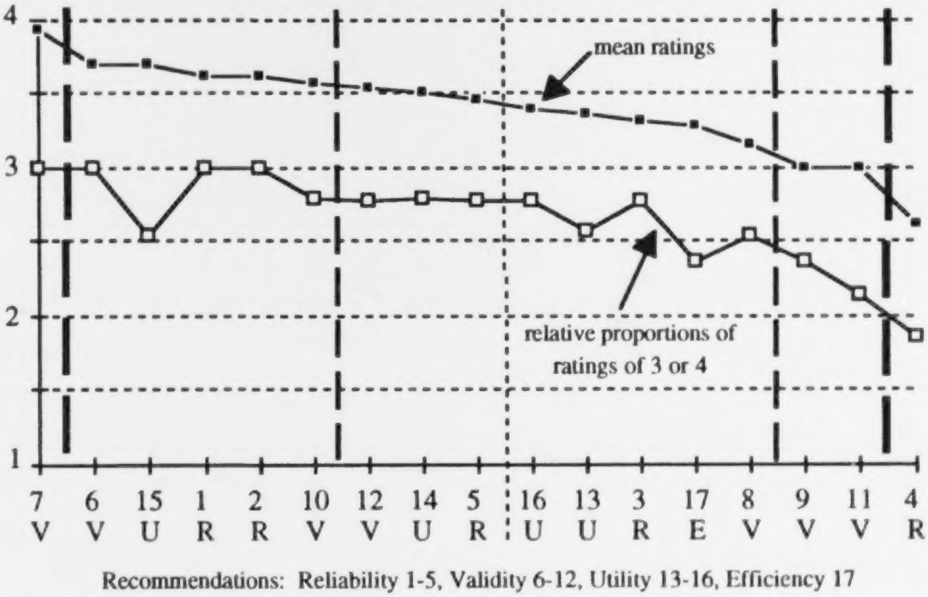
17. Making maximum use of assessment time and effort.

## Ratings of Importance of the Recommendations

As noted in Figure 2 above, reviewers' ratings of the importance of the recommendations were scaled from 4 for *very important* to 1 for *not at all important*. One of the 15 reviewers who responded to the other structured scales did not respond to the importance scale, and another responded to it for only seven of the recommendations. Mean ratings gave an indication of perceived importance, but the breakdowns of rating frequencies were also informative, and these are discussed as well. Of the 136 pairs of means 17 achieved significance at the .05 level and 12 at the .01 level. Recommendation 7 was rated as most important ($M = 3.9$): it was rated significantly more important than recommendations 4, 8, 9, and 11 ($p<.01$), and than 3, 13, and 17 ($p<.05$). Recommendation 4 was clearly rated less important than the others ($M = 2.6$): $p<.01$ for most. Recommendations 6, 15, 1, 2, and 10 were rated higher than recommendation 4 ($p<.01$), and than 9 and 11 ($p<.05$). The recommendations were then ordered according to mean rating of importance; this is depicted in Figure 4.

The groupings of recommendations that reflect the apparent differences in importance are noted in Figure 4 by broken vertical lines, the heavier lines separating the most clearly differentiated recommendations, namely 7 and 4. The means of the six higher-rated recommendations (7, 6, 15, 1, 2, 10) all exceeded 3.5. The majority of reviewers rated all six as *very important* (8 or more of the reviewers), and no reviewers rated them as *not at all important*. Except for recommendation 15, no reviewers rated these six recommendations below the highest two categories on the 4-point scale (in Figure 4 this is reflected by the dip in the graph for recommendation 15 of *relative proportions of ratings of 3 or 4*; two reviewers rated this in the second lowest category, a 2). It is difficult to distinguish clearly between this group of recommendations and the next three recommendations (12, 14, 5), which all have mean ratings near 3.5. These also received very few ratings in the lower two categories of the scale.

**Figure 4**. Recommendations Ordered According to Reviewers' Mean Ratings of Importance (scale of 4 to 1); Also Shows Proportions of Reviewers Rating Each Recommendation as *Very important* (4) or the Next Scale Category (3)



Note. The proportions of reviewers selecting *very important* (4) or next category on the scale (3) were converted to a scale from 0%, indicated by the scale point 1, to 100%, indicated by the scale point 3, so that these could be displayed on the same chart as the means.

The remaining eight recommendations received mean ratings of 3.4 or lower. This demarcation is noted in Figure 4 by a lighter dotted vertical line. All had more ratings in the lower two categories of the scale, indicating that several reviewers considered them less important. This was clearly apparent from the graph of proportions in Figure 4 for recommendations 13, 17, 8, 9, 11, and 4, where from two to five reviewers rated them in the lowest two categories (a 1 or 2), but less apparent for recommendations 16 and 3.

The order of importance of the 17 recommendations, and that nine of them were more highly rated than the other eight, form the basis of the presentations and discussions of recommendations that follow. Directly below are summaries of the results for the four features of instructional delivery. In the next section on implementation of the recommendations, the recommendations were also grouped according to their importance, and the results of the review are discussed in detail for each group of recommendations.

## Responses to Feature 1--Level of Specificity for Instruction

The first of the four features addressed the level of specificity of instruction the reviewers thought appropriate for each of the 17 recommendations. This was whether instruction should be *theoretically based and developed from general principles* (4) or *specific and providing classroom-oriented prescriptions* (1), with two response categories given between these points (see Figure 2 above). Some reviewers checked two or more categories on the scale, such as both the first and the last, although the number of multiple responses was not large. This is not unreasonable as it is possible that both theoretically based and specific prescriptions are appropriate for instruction in a topic related to a particular recommendation. If a reviewer gave two or more responses, the mean value of these responses was entered.

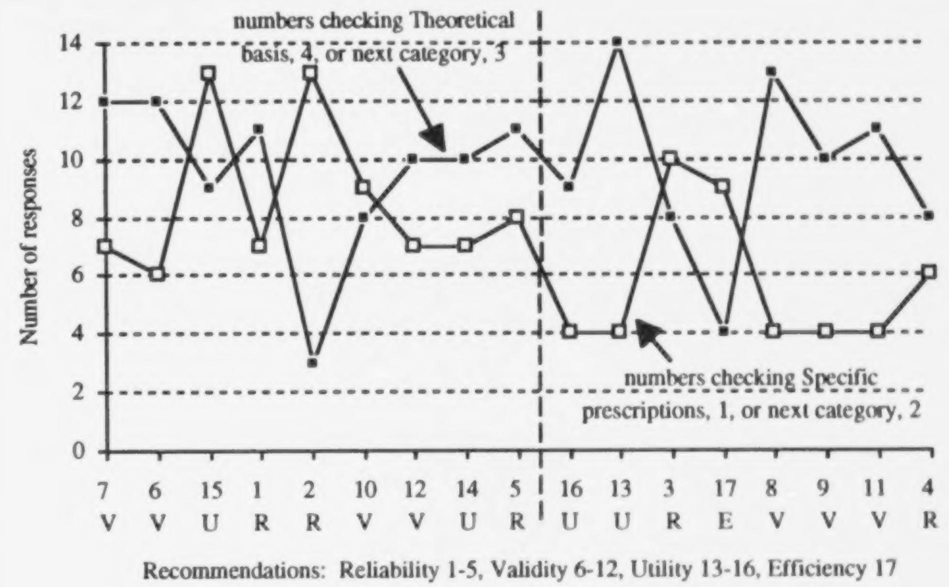
A scale mean above 2.5 suggests that on the whole reviewers thought the instruction for the recommendation should be more theoretically based, and a mean below 2.5 suggests instruction should emphasize specific classroom prescriptions. The mean for recommendation 2 was 1.9, which was significantly below 2.5 ($p<.05$), suggesting that instruction could be more specific and prescriptive. The means for recommendations 8 and 11 were 3.0 and 3.2 respectively, which were significantly above 2.5 ($p<.05$), whereas those for recommendations 9 and 13 were 3.2 and 3.3 respectively, which were significantly above 2.5 ($p<.01$). This suggests that instruction for these recommendations could be more theoretically based. Means for the remaining recommendations did not differ significantly from 2.5 (they ranged from 2.2 to 2.9), and instruction for these may combine theoretical developments with specific prescriptions, but not emphasize either.

This approach to interpreting the results masks the possiblity that some reviewers strongly endorsed theoretically based instruction whereas others endorsed specific prescriptions. Therefore, the next step was to combine the numbers of responses for scale points 4 and 3, which represented instruction that should be primarily theoretically based, and those for scale points 2 and 1, which represented instruction in specific classroom prescriptions. These results are depicted as a graph in Figure 5 with the recommendations ordered according to their importance.

From this, it appeared that for recommendations 7 and 6 many of the reviewers (12 responses) thought instruction could be more theoretically based, although the means were not significantly above 2.5. For recommendations 15 and 2, reviewers in the main (13 responses) believed instruction should be specific and prescriptive, and for recommendation 2 this was supported by the significance test of the mean. For recommendation 1 more reviewers (11 responses) thought instruction should be theoretical than specific prescriptions (7), although the mean was not significantly different from 2.5. For recommendation 10 reviewers were fairly evenly split as to whether instruction should be theoretical or specific. For recommendations 12, 14, and 5 more reviewers (10-11 responses) appeared to support theoretical instruction, although the means did not differ significantly from 2.5.

For six of the remaining eight recommendations (those rated as less important), more reviewers (8-14 responses) thought instruction should be theoretically based: for recommendations 16, 13, 8, 9, 11, and 4. Of these, for recommendations 13, 8, 9, and 11 the means were significantly higher than 2.5. For recommendations 3 and 17 specific instruction was favoured by more reviewers (9-10 responses), but only slightly and the means did not differ significantly from 2.5.

**Figure 5**. Numbers of Reviewer Responses to Levels of Specificity for Instruction in Assessment, on the Scale *Theoretical basis* (4) to *Specific prescriptions* (1) ($n$ = 15)



Note. The numbers of reviewer responses selecting *theoretical basis* (4) or next scale category (3) were combined, and numbers selecting *specific prescriptions* (1) or next scale category (2) were combined.

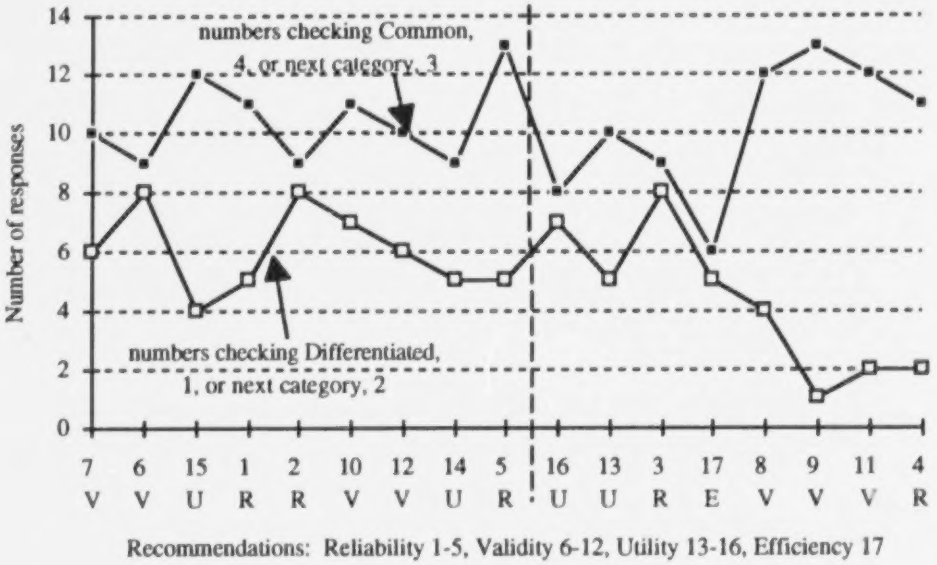
## Responses to Feature 2--Common or Differentiated Instruction

The second of the four features was whether instruction should be common for all prospective teachers or differentiated for particular subgroups, or if certain aspects should be taught in a common setting and others not. The scale had four categories for responses, which were scaled from 4 to 1 with end points being *common program for all students* and *differentiated for particular groups* (see Figure 2 above). Some reviewers checked more than one category on the scale, such as common program as well as differentiated. As with Feature 1, this is reasonable since some topics pertinent to a recommendation could be taught in a setting common for all students, but others ought to be differentiated for particular groups, such as those in elementary education programs versus secondary programs. If a reviewer gave more than one response for a recommendation the mean of these responses was entered, but few reviewers did so.

Results are presented similarly to those for Feature 1. Means above 2.5 suggest instruction could be common for most education students, whereas for those below 2.5 instruction should be differentiated for some groups of students. The means for recommendations 4, 9, and 15 ranged from 3.3 to 3.6 and were significantly above 2.5 ($p$<.01). Those for recommendations 5, 8, and 11 ranged from 3.0 to 3.3 and were significantly above 2.5 ($p$<.05). This supports the interpretation that instruction for these recommendations should be generally common for students. No mean was found to be significantly below 2.5; in fact, none was below 2.5. This suggests that for the remaining recommendations reviewers were not definitive as to whether instruction should be common or differentiated. But in the main reviewers favoured common instruction over differentiating it for various subgroups.

For the second analysis, the numbers of responses were combined for scale points 4 and 3, which reflected a preference for common instruction, and for 2 and 1, which reflected a differentiated preference. The recommendations were ordered according to the importance of the recommendations reported earlier, and these combined frequencies depicted as a graph in Figure 6.

**Figure 6**. Numbers of Reviewer Responses to Level of Program Differentiation for Instruction in Assessment, on the Scale *Common for all students* (4) to *Differentiated for particular groups* (1)



Note. The numbers of reviewer responses selecting *common program* (4) or next scale category (3) were combined, and numbers selecting *differentiated* (1) or next scale category (2) were combined.

The dotted line on Figure 6 separates the nine recommendations rated higher in importance from the others. It appears that instruction for recommendations 7, 6, and 2 could be either common or differentiated (not significantly different), and only several more reviewers chose common instruction as chose differentiated instruction (9-10 responses versus 6-8). But for recommendations 15 and 5 the means were significantly higher than 2.5, and more reviewers (12-13 responses) favoured common instruction over differentiated (4-5 responses). Means were not significantly different from 2.5 for the other four recommendations that were placed in the more important group, recommendations 1, 10, 12, and 14 (to the left of the dotted line in Figure 6). But reviewers appeared to favour common instruction for these recommendations, with from 9 to 11 responses in categories 3 or 4, versus from 5 to 7 in categories 1 or 2.

Of the eight recommendations awarded lower importance, four obtained means significantly above 2.5: recommendations 8, 9, 11, and 4. These received from 11 to 13 responses in categories 3 or 4, clearly supporting common instruction. The remaining recommendations (16, 13, 3, 17), whose means were not significantly different from 2.5, received from 8 to 10 responses of 3 or 4, and from 5 to 8 responses of 1 or 2. This suggests that reviewers were divided on common versus differentiated instruction for these four recommendations.

## Responses to Feature 3--Method of Delivery of Instruction

For the third feature there were four response categories: part of a distinct course, short course, seminar or lecture, and part of another education course (see Figure 2 above). Reviewers were also asked to comment on the method of instruction including timing and location of this in a student's program. Significant differences in frequencies for the four response categories were obtained at the .05 level for recommendations 5, 6, 7, and 8, indicating that there were differing numbers of responses for some of the four categories. Recommendations 5, 6, and 7 were rated as more important by the reviewers (recommendations are summarized in Figure 3 and the order of importance is given in Figure 4 above). For recommendations 5 and 7 the response frequencies suggested more support for the delivery method *part of a course on pedagogy* (11-12 responses), which received by far the highest frequencies of response, followed by *part of one course* (6 responses). Reviewers tended to believe that instruction for these recommendations could be part of instruction in pedagogy, although some believed that a separate course is preferred. For recommendations 6 and 8 the same two categories were favoured, but in this case with nearly equal frequencies (6-8 responses). Few reviewers (1- 3 responses) thought short courses or seminars were appropriate for these four recommendations. This implies that for these recommendations in particular, instruction should be either as part of instruction in pedagogy or that it should be as a course devoted to assessment.

There appeared to be a general finding that the two categories favoured for these four recommendations were more endorsed for all recommendations: the two categories received total response frequencies for the 17 recommendations of 106 and 124 respectively, as opposed to the short courses and seminars receiving 44 and 46 respectively. This suggests that the reviewers generally favoured instruction for the recommendations as part of instruction in pedagogy or in a course devoted to assessment, and that short courses and seminars would not be appropriate. However, there were recommendations for which the reviewers did not favour one method over the others.

For the six additional recommendations that were rated as more important, 1, 2, 10, 12, 14, and 15, it was not as clear what method of delivery the reviewers favoured. Significance was not achieved, although for recommendations 1, 2, and 15 the two categories, part of one course and part of a course on pedagogy, were endorsed with almost equal frequency, 6 to 8 responses, and more frequently than the other two categories by 3 to 6 responses. For recommendation 12 the category of part of a course on pedagogy was selected by slightly more reviewers, but the frequencies for the other three categories were similar. Recommendations 10 and 14 received similar frequencies for three categories, with the seminar category receiving somewhat fewer responses.

The remaining seven recommendations received lower ratings of importance (recommendation 8 also was rated less important but it was significant so it was addressed above). For recommendations 3, 16, and 17 the category, part of a course on pedagogy, received from 3 to 6 more responses than did the other three categories. For recommendations 9 and 11 the category, part of one course, received from 3 to 4 more responses than did the other three categories. For recommendation 13 the two categories, part of one course and part of a course on pedagogy, received similar numbers of responses and more than the other two categories. The recommendation rated lowest in importance, 4, received similar numbers of responses for all four categories.

## Responses to Feature 4--Nature of Instruction

For this last feature reviewers were asked to indicate what would be the best approach to the nature of the instruction for each recommendation: lecture/discussion, laboratory work, clinical practice, and practical in-school experience (see Figure 2 above). Reviewers were asked to select one or more of these four, and to comment. As with the third feature, frequencies of responses were compared across the four categories for each recommendation, and of the 17 only recommendations 4 and 8 achieved significance ($p<.01$), indicating that there were differing numbers of responses for various categories. Both of these recommendations were rated as less important by the reviewers (for the order of importance see Figure 4 above), so are only commented on briefly. For both recommendations reviewers clearly favoured the categories *lecture/discussion* (10-11 responses) and *laboratory work* (6-9 responses).

The responses for each category, when totaled for all recommendations, favoured lecture and discussion over the other three approaches ($p<.01$), with 161 responses compared to total frequencies in the 80's and 90's for the other three. This was supported in that for all recommendations except recommendation 3 the first category received 25% or more of the responses. This suggests that the reviewers generally supported instruction using lecture and discussion for the recommendations, but that the other three approaches were more or less appropriate depending on the recommendation.

Significance was not obtained for any of the nine recommendations rated higher in importance by the reviewers. For recommendation 7 the frequencies ranged from 6 to 10, and for recommendation 6 from 4 to 8, suggesting similar support for all four categories of instruction: instruction could include lecture and discussion, laboratory work, clinical practice, and in-school experience. However, more responses favoured lecture and discussion for recommendations 1, 5, 10, 12, and 15, where in each case this category exceeded the other three by 5 or more responses. The other three categories received similar numbers of responses for recommendations 1, 5, and 10, and the category in-school experience received 3 to 4 more responses than the other two for recommendation 15. For recommendations 2 and 14 the responses to the lecture and discussion category exceeded those to the other categories by 2 or 3 responses. Clinical experience and in-school experience received nearly as many responses for recommendation 2, whereas laboratory work received nearly as many for recommendation 14.

The remaining six recommendations had received lower ratings of importance (recommendations 4 and 8 also were rated less important but they were addressed above). There was a similar pattern across categories for recommendations 3, 16, and 17, where the numbers of responses deviated from one another by 3 or less. This suggests that all approaches were considered similarly appropriate. For recommendations 9 and 13 the category, lecture and discussion, received from 5 to 7 more responses than did the other three categories. Finally, for recommendation 11, the two categories lecture and discussion and laboratory work seemed to be supported more than the other two.

## Implementation of the Recommendations

Reviewers' ratings of importance indicated which recommendations were considered more important (Figure 4 above) and therefore should form the focus of teacher preparation in classroom assessment. Nine recommendations appeared to be accorded more importance than the other eight. These nine were further divided into three groups and are discussed below in these groups. For convenience, summary statements for the recommendations are given in this order in Figure 7. More detailed statements of the recommendations are given in the context of the discussion and also in Appendix D. The discussion includes interpretation of the perceived importance of each recommendation, but focuses more on how instruction should be implemented.

**Figure 7.** Summary Statements of the 17 Recommendations Ordered and Grouped According to Reviewers' Ratings of Importance

| | **The nine recommendations rated higher in importance** |
|---|---|
| **7.** | Assessment that supports and informs good instruction. |
| **6.** | Increasing assessment of important, higher level cognitive learnings. |
| **15.** | Communicating the results and interpretations of assessments. |
| **1.** | The importance of reliability for high-stakes assessments. |
| **2.** | Practical ways of improving reliability in the classroom. |
| **10.** | Making assessment results reflect the focus and purpose of the assessment. |
| **12.** | Removing bias and prejudice in classroom assessments. |
| **14.** | Obtaining the scores necessary to use the assessment results properly. |
| **5.** | Reliability of subjective forms of assessment. |
| | **The eight recommendations rated lower in importance** |
| **16.** | Controlling the effects of subjectivity in observing and marking. |
| **13.** | Making value interpretations of assessment results using referencing procedures. |
| **3.** | Increasing reliability through multiple quality assessments. |
| **17.** | Making maximum use of assessment time and effort. |
| **8.** | Assessment for various purposes including instructional diagnostics. |
| **9.** | Empirical checks of validity of the skills covered in classroom assessments. |
| **11.** | Confirming the results of assessments by comparison with other assessments. |
| **4.** | Numerical procedures for estimating reliability in classroom assessments. |

## Recommendations 7, 6, and 15

Of all 17 recommendations the 7th was given the highest rating of importance. It received a mean rating of 3.9, on a scale of 1 to 4, which was significantly higher than seven of the others. Recommendation 6 followed in importance with a mean rating of 3.7, which was rated significantly higher than three others. Both deal with validity, and emphasize that teachers must assess in appropriate ways the skills and learnings that are most important for their students and for the school program. Recommendation 15 was also considered important, and, with a mean of 3.7, it was rated as significantly more important than three others. It falls under the heading of utility and reflects a practical concern of educators for the need to clearly communicate the results of assessments to students, parents, and others. In detail, the three recommendations are:

> **Validity 7. Assessment that supports and informs good instruction.** Teachers must have training and experience in preparing and using assessment procedures that support the most important learning in classrooms, those that should be emphasized in class and on the assessments. Teachers must know how to

develop and use paper-and-pencil techniques, but more importantly, they should include systematic observation and applied performance assessment.

**Validity 6. Increasing assessment of important, higher level cognitive learnings.** Classroom assessments must provide information on the learning that is most important for our students, which includes more complex and higher-level thinking. Teachers must have the knowledge and skills to design and conduct classroom assessments that are most appropriate to this kind of outcome.

**Utility 15. Communicating the results and interpretations of assessments.** Various audiences, with different expectations and different abilities, must understand the results. It is necessary to simplify interpretations of assessment results in many cases. Teachers must be able to communicate assessment results in variety of ways and to different audiences. This would include numerical results, summary statistics, grades, anecdotal reports, diagnoses, and affective evaluations.

Recommendation 15 does not include interpretation of assessment results in the narrow sense of criterion- or norm-referencing, which is embodied in recommendation 13 (which was rated somewhat lower in importance than 15 although not significantly so). Recommendation 13 was not highly rated--perhaps because score referencing is not a clear concept for most educators, and it may be more aligned with norm-referenced testing and the external setting of standards than with classroom assessments.

Clearly, the reviewers thought that these three recommendations are important for teacher education. If the goals of classroom instruction are in keeping with present curriculum expectations, which highlight higher-level thinking (e.g., Yager, 1989), the two validity recommendations can be taken in conjunction with one another. Recommendation 15 also can be taken in conjunction with these validity recommendations. As a group the three imply that teachers ought to be able to identify and assess the most important learning objectives, particularly higher level thinking skills. And this ought to be in such a way that the results can be communicated clearly to students and to others, and used by the teachers to inform their instruction.

For the three recommendations in this group, reviewers' responses to the level of specificity of instruction they thought appropriate were not significantly different from a point midway between the extremes of theoretical basis to specific prescriptions. For the validity recommendations there was support for instruction in the theoretical basis for assessing higher level thinking and using this to guide teaching, but also for specific prescriptions for teacher practice as well. Figure 5 shows that 12 responses supported theoretical instruction, but that there were 6 responses for recommendation 6, and 7 for recommendation 7, that supported instruction in specific classroom assessment prescriptions. One reviewer noted that specific prescriptions may be difficult for higher-level skills, since by their very nature these skills do not lend themselves to a set of best procedures, although general principles and guidelines can be given (e.g., Nickerson, 1989a, b; Norris, 1989; Wolf, Bixby, Glenn, & Gardner, 1991). For recommendation 15 more reviewers, 13, endorsed specific classroom prescriptions than theoretical development, 9. But, as one reviewer stated, teachers "need to understand the meanings of assessment results themselves before they can interpret them to others", suggesting that although teachers should have specific preparation in communicating assessment results, this should be supported with some theoretical background. Since these are important recommendations to teacher preparation it seems that there should be considerable theoretical development as well as actual training in specific classroom applications.

For all three recommendations more reviewers endorsed a program of instruction common for all prospective teachers than ones differentiated for particular groups. However, on a scale of 1 to 4, the means for recommendations 6 and 7 were only slightly above 2.5, and not significantly so. This suggests minimal support for common instruction over that for differentiated instruction to all students in the areas of assessing higher level thinking, and using this to guide teaching. For recommendation 7, there were 10 responses that supported common programs and 6 that supported differentiated programs, but for recommendation 6, the number of responses supporting common programs, 9, was almost equal to those for differentiated programs, 8 (see Figure 6). Reviewers did note, however, that topics related to these recommendations could be taught in a common program for all students but that some differentiation is necessary, particularly for teachers specializing in various subject areas.

For recommendation 15 common instruction was favoured over differentiated: the mean was 3.3, which significantly exceeded 2.5 ($p<.01$). The majority of responses, 12, favoured common instruction, as opposed to 4 responses for differentiated instruction. Clearly, the reviewers thought that prospective teachers could receive common instruction in the communication of assessment results, but several comments indicated that some adjustment for elementary and secondary programs may be desirable. This differentiation would appear reasonable since reporting procedures, for example, are often much different between the lower and higher grade levels.

For recommendations 6 and 7 there were significant differences ($p<.05$) in frequencies of responses across the four methods identified for delivery of instruction. For recommendation 7 there was more support for the two delivery methods part of one course and part of a course on pedagogy, the latter receiving by far the highest frequency of response, 12. For recommendation 6 these two delivery methods were favoured as well, but with nearly equal frequencies. Reviewers were split as to whether these recommendations should be covered in a separate course or part of courses on pedagogy, but generally agreed that short courses and seminars were not applicable. The nature of these recommendations probably requires that there be longer, more extensive instruction, and integration of the topics with pedagogy. This suggests that instruction in assessing higher cognitive learning (recommendation 6) and in assessing to inform appropriate instruction (recommendation 7) should be part of a distinct course as well as part of a course on pedagogy. For recommendation 15 there were no significant differences noted, and support for the four categories of delivery could not be clearly differentiated, although higher frequencies occurred for the same two delivery methods as for recommendations 6 and 7. The communication of assessment results referred to in recommendation 15 probably could be capsulated in smaller units, such as short courses or seminars, more readily than topics for recommendations 6 or 7, so instruction may be possible to be delivered in various ways, including smaller units as well as part of a distinct course.

Finally, for these recommendations, reviewers gave some support for all four approaches to the nature of instruction: lecture, laboratory work, clinical practice, and in-school experience, but perhaps with lower support for laboratory work. Frequencies of responses were similar for the four approaches, ranging from 4 to 8 for recommendation 6 and from 6 to 10 for recommendation 7. Somewhat more support was indicated for the lecture format for recommendation 15, 12 responses, although this too was not significant. For all three recommendations total numbers of responses were considerably higher than one per reviewer suggesting that more than one approach to instruction were considered appropriate.

The comment was made that there was not a strong research base on which to develop instruction related to recommendation 15, communicating assessment results.

Certainly there are plenty of suggestions based on analysis and professional experience (e.g., Gronlund & Linn, 1990; Hills, 1981; Taylor, 1979; Terwilliger, 1989), and research is beginning but still mostly at the high school level (e.g., Frary, Cross, & Weber, 1992; Friedman & Manley, 1991; MacRury, 1988; Stiggins, Frisbie, & Griswold, 1989;). Needs of elementary teachers have not been addressed to any great extent--this is just starting (e.g., Brookhart, 1992; Manke & Loyd, 1991; Nava & Loyd, 1992), and yet we know their situation is different (in fact, for example, anecdotal reporting has become a mainstay at the lower grade levels; e.g., Schulz, 1993).

In summary, the review indicated that instruction related to assessing higher-level cognitive learning, assessing to inform appropriate instruction, and communicating assessment results is important. Reviewers thought that there ought to be both theoretical development of these topics and specific classroom prescriptions for teacher practice, although there may have been a slight emphasis on the theoretical for the assessment of important learnings and how assessment might inform teaching, and on specific prescriptions for the communication of results. The delivery of instruction could vary. While more reviewers thought a program common to all prospective teachers would be appropriate for these recommendations, this was considerably more pronounced for the communication of assessment results. Some reviewers thought differentiation would be desirable for part of what might be covered under the two validity recommendations. Reviewers clearly favoured instruction for the recommendations to be part of other educational coursework or in a separate course, although slightly less so for the communication of assessment results. Some reviewers endorsed lecture and discussion, laboratory work, clinical practice, and practical in-school experience as approaches instruction for these recommendations, suggesting all should be seriously considered.

## Recommendations 1, 2, and 10

Following recommendation 15 in importance were two concerned with reliability, recommendations 1 and 2, and one with validity, recommendation 10. All three were rated significantly more important than the three recommendations rated lowest in importance, and all had similar mean ratings of 3.6. Recommendation 1 is based on the principle that for assessment purposes with significant personal and other consequences (i.e., high-stakes tests) high reliability of the measures is more important; this relates to the consequential aspects of evaluation, and hence is important to validity. Recommendation 2 deals with the practical ways of enhancing reliability of assessment in a classroom setting. Recommendation 10 deals with an aspect of validity that is somewhat different from recommendations 6 and 7: it refers to the structural aspects of validity and how assessments are designed to produce the kinds of results appropriate to the purposes of the assessment. Both recommendations 1 and 10 emphasize how the purposes of an assessment impact on its design. The recommendations are:

> **Reliability 1. The importance of reliability for high-stakes assessments.** The importance of reliability is directly dependent on the consequences of the assessment. Therefore, teachers must be able to identify the main purpose of an assessment and its implications for the student, both the personal and social consequences (such as in high-stakes assessments).
>
> **Reliability 2. Practical ways of improving reliability in the classroom.** The reliability of assessment information can be enhanced in two general ways which are practical in the classroom: by making the assessment procedures more explicit and systematic, and by increasing the amount of high-quality information gathered. Teachers must know how these apply to particular classroom assessment practices.

**Validity 10. Making assessment results reflect the focus and purpose of the assessment.** Teachers must be trained to obtain scores from assessments corresponding to the nature and purpose of the assessment. This involves identifying clearly in advance of producing the assessment the approximate importance, and hence the weighting, of various areas of content and levels of skills (essentially producing a table of specifications for the assessment).

Recommendation 2 is related to recommendation 3, which received somewhat lower ratings (but not significantly lower than ratings for 1 and 2). Recommendation 3 refers to the need for multiple assessments over a period of time to ensure that evaluation is based on stable measurements. Recommendation 10 is related to recommendation 14 under utility which was rated almost as highly, and to recommendation 8 in validity which was rated somewhat lower (although not significantly so). Recommendation 10 argues that teachers must produce assessments and scoring procedures which yield scores that correspond to the purposes of the assessment, and that do not subvert the intent of the assessment by such things as inappropriate emphasis or weighting. In a broad sense, this relates to recommendation 7 since an assessment must be able to produce the types of information necessary to meaningfully inform instruction. Recommendation 14 refers to the discriminating power of scores, whereas 8 refers to the notions of subtest score interpretation and diagnostic profiles.

Reviewers were split as to whether instruction for recommendation 1 should be primarily theoretically based or specific and prescriptive. The mean of responses was 2.7, which was only slightly and not significantly above the scale middle point of 2.5, suggesting only minimal support for theoretical development over specific prescriptions, and that both are probably desirable. Nearly as many responses endorsed specific prescriptions, 7, as endorsed theoretical development, 9 (see Figure 5). The recommendation deals with enhancing reliability for high-stakes assessments, and it appears appropriate to develop this both theoretically and practically. However, for reliability recommendation 2, reviewers supported specific prescriptions over theoretical development: the mean of 1.9 was significantly below 2.5 ($p<.05$), and responses favoured specific prescriptions 13 to 3. This recommendation deals with practical ways of improving reliability of assessment, so these results also appear to be appropriate.

The responses to recommendation 10 (validity) were less clearly interpretable. The mean was 2.5 with near equal numbers of responses endorsing theoretical development as specific prescriptions. This was confirmed in their comments: both were desirable for instruction in this area. Perhaps this recommendation is ambiguous itself, but it refers to what seems to be a fairly practical matter of ensuring that the assessment and the scoring reflect the purpose the results are to serve, and that the assessment procedures are in keeping with the overall intent. Assuring that weighting assessment components and forming grades in concert with instructional intent is often a more complex, technical problem (e.g., Thayer, 1991) but the overall process usually begins with setting a table of specifications, followed by devising items and scoring procedures that are in keeping with the assessment's purpose (e.g., summative, formative, diagnostic) and the intended importance of various topics and aspects. It may be possible to instruct teachers in the general principles of designing assessments so the types of results are in keeping with the purposes, although this is usually not treated as a theoretical matter and certain specific procedures are almost ubiquitously recommended by measurement experts, such as producing tables of specifications and identifying behavioral outcomes (e.g., Gronlund & Linn, 1990; Mehrens & Lehmann, 1991; Nitko, 1983; Thorndike, Cunningham, Thorndike, & Hagen, 1991; see also Table 8 above). The evidence to date suggests that teachers do not use these procedures systematically for assessments, however (e.g., Marso & Pigge, 1992; the case studies). Further, the evidence is fairly clear that teachers' assessments do not necessarily reflect the intent of the curriculum, and often not that of their own teaching: compelling examples of this are the scarcity of higher-level thinking questions on tests (e.g., Fleming & Chambers, 1983; Haertel, 1986; Stiggins, 1986a; Stiggins, Griswold, & Wikelund, 1989; the case studies) and the lack of assessments of process and other skills (e.g., Wiggins, 1989b; the case studies).

For none of the three recommendations did reviewers clearly favour either a common program of instruction or one differentiated for subgroups of education students. However, for recommendations 1 and 10, more reviewers selected a common instructional program, 10 and 11 responses respectively, than selected a differentiated one, 5 and 7 respectively (Figure 6). Mean responses were 2.9 and 3.0, which were not significantly different from 2.5. Common instruction should be acceptable for most topics related to these recommendations since many of the issues are similar for elementary and secondary classrooms, but reviewers felt that differentiation was desirable for various subject areas and specialties: topics such as designing systematic assessment procedures and assuring that they yield stable results. The problems of producing assessment scores which are in keeping with the intent of the assessment may be taught to groups in common, but one could expect that evaluation purposes and practices differ considerably from the higher to the lower grades. However, this is an issue for both tests and other forms of assessments, and applicable at all levels.

For recommendation 2 the numbers of responses were nearly equal, and the mean response was 2.6, suggesting that the instruction could be either common or differentiated. Here, as well, some topics may be taught to common groups of education students whereas others may be better taught to groups differentiated for different grade levels of students, such as the development of observation schedules and rating procedures, and could incorporate actual curriculum content. Reviewer suggestions included providing students in common with the basic principles and ideas, and providing specific instruction to each of elementary and secondary education students.

For these three recommendations no significant differences were noted in number of responses to the methods of delivery of the instruction. For recommendations 1 and 2, though, there appeared to be greater preference for instruction to be part of one course or part of a course on pedagogy, rather than as mini-courses or seminars and lectures. Topics for these recommendations may require greater integration with other measurement topics or with pedagogy, although this would be expected for recommendation 10 as well since it to deals with the purpose of the assessment and how this defines the structure of the assessment. One suggestion was that the general principles can be developed in a course, and that these be supported in courses on pedagogy. For recommendation 10 the methods by which this should be taught could include separate course, short course, and part of course on pedagogy, with less support for seminar: responses ranged from 2 to 7 for the four choices. The problems of producing assessment scores which are in keeping with the intent of the assessment may be taught as a unique topic in a short session, or as part of instruction in assessment or pedagogy.

Reviewers did not clearly identify one approach to instruction as being preferable to the other three identified, although for the three recommendations there were from 10 to 12 responses favouring lecture and discussion and only 7 or fewer favouring the other three approaches. Reviewers tended to select more than one approach since total number of responses exceeded the number of reviewers. No other approach seemed to be preferred, however, and all three received some endorsements. It appears from this that instruction should include lecture and discussion, but may well incorporate laboratory work, practice, and in-school experience to address these recommendations. Some reviewers indicated the value of laboratory work to these topics. Reviewers who were practicing teachers and curriculum educators in particular felt that there should be specific instruction in recommendations 1 and 2, and this should correspond to the grade levels of the children and to curricular areas. They also were more likely to suggest that much of this be taught as part of courses in pedagogy.

## Recommendations 12, 14, and 5

These three recommendations were rated as slightly less important than the recommendations discussed above, although not significantly so. Their mean ratings approached 3.5, which were significantly higher ($p<.01$) than the rating for recommendation 4 only. Recommendation 12 addresses an aspect of validity that is related to the validity recommendations discussed above, but it focuses on the control of bias and prejudice in assessments. Utility recommendation 14 deals with designing assessments so that they yield scores that discriminate performances at points on the scale where this is most necessary, such as at pass/fail points. This is related to recommendation 10 but focuses more directly on the types of scores obtained and the error at particular points on the scale. Recommendation 5 deals specifically with reliability of subjective assessments. They are:

> **Validity 12. Removing bias and prejudice in classroom assessments.** Direct deleterious effects on the assessment results of particular groups of students due to their differences in experience which makes them perform below their actual skill level is an issue of fairness or equity. The kind of bias or prejudice that has a negative reflection on particular subgroups may over the long run have a negative impact on these subgroups. Teachers must recognize and control biases of both types in the materials they use, expectations they have of students, and how they relate to students.
>
> **Utility 14. Obtaining the scores necessary to use the assessment results properly.** Appropriate kinds of assessment results and interpretations must be obtainable from the assessment. Teachers must learn how to design the assessment so that the appropriate information is forthcoming. The assessment must produce discrimination of scores at useful points on a scale if the scale is to be converted into grades or mastery-nonmastery categories.
>
> **Reliability 5. Reliability of subjective forms of assessment.** Teachers must understand that subjective forms of assessment are prone to unreliability and to personal bias. They must also know practical procedures which can ameliorate these problems.

Recommendation 12 refers to a consequential aspect of validity, and argues that teachers must control for bias and prejudice in their assessments. This can subsume recommendation 5 which deals with reliability of subjective forms of assessment, but 12 includes a much greater range of potential biases. Recommendation 5 is restricted to the vagaries directly associated with teacher subjectivity in marking student work, and this is treated as a component of reliability. Recommendation 12 can subsume also the problem addressed under utility recommendation 16, that of subjectivity in observing and marking a broad variety of assessment procedures including direct assessment of student behaviours and performances. However, the notions of bias and prejudice in recommendation 12 refer to all aspects of the assessment, not just that of scorer subjectivity--biases that may be even more insidious: this addresses the focus of the assessment, the type of assessment instrument and procedures used, the selection of content to be tested and material that appears in the assessment, and the language and wording of the assessment, as well as what are considered good student responses and behaviours. Recommendations 12, 5, and 16 are related, but recommendation 16 received a slightly lower rating of importance (although not significantly lower) so was grouped for discussion in the section that follows.

For recommendations 12, 14, and 5 reviewers appeared to favour instruction of a theoretical nature, but only mildly so: the means of 2.7 to 2.8 were not significantly different from a middle point of 2.5. There were from 10 to 11 responses which supported a theoretical basis and 7 to 8 responses which supported specific prescriptions (Figure 5). Instruction related to recommendation 12 should provide an understanding of how bias and prejudice can permeate assessment practice, and it could identify a theoretically based rationale for a set of principles to control bias but this should include sufficient practice in applying these to actual assessment procedures. Part of this instruction could include enhancing reliability of subjective assessments, although this may by more appropriately developed as a topic in its own right, recommendation 5, and include both theoretical and practical development; in fact, this was suggested in reviewer comments. Since it addresses reliability of subjective assessments, this appears both reasonable and possible. The results for recommendation 14 suggest that there ought to both theoretical background on how to design assessments that produce scores which are sensitive where necessary, and practical work on what this means in the classroom.

Reviewers generally supported instruction for these three recommendations that would be common for students. Mean responses for recommendations 12, 14, and 5 were 3.0, 2.9, and 3.2 respectively: all were above the middle point of 2.5, although only for recommendation 5 was this significant ($p<.05$). Recommendation 12 received 10 responses, and recommendation 14 received 9 responses, endorsing common programming as compared to 6 and 5 responses respectively that endorsed differentiation (Figure 6). This suggests that much of what may apply to issues of bias and prejudice in assessment could be treated in a common setting, with perhaps some differentiation. Instruction related to recommendation 14 could also be dealt with in a common setting, and some issues are similar for all levels of the school system, but there may be need for some differentiation since the level of expectation varies across subject areas, and certainly with grade levels. Some issues are more relevant to certain groups of teachers: for example, assessing mastery probably would be more appropriate to elementary programs and to vocational training, whereas grading and pass/fail decisions become more of an issue at the higher grade levels and in academic subject areas. In contrast, recommendation 5 received 13 responses endorsing common programming, versus 5 for differentiated. Subjective forms of assessment occur at all levels of the school system and in all subject areas, although perhaps more in some areas than others, and could probably be dealt with effectively for all students in common. There is some argument, though, that differentiation may be valuable to provide instruction on specific, practical procedures relevant to selected subject areas (e.g., language arts) and to students grouped according to grade level (e.g., early years, middle years, senior years).

Reviewers did not clearly favour a particular method of instructional delivery for recommendations 12 and 14: response frequencies did not differ significantly for the four categories. But for recommendation 5 reviewers did significantly differentiate among the categories, and appeared to favour teaching these aspects as part of course on pedagogy. The results for recommendation 12 suggest that issues of bias and prejudice in assessment could be treated in a short, discrete session, or as part of a course on assessment or pedagogy, although there were more responses favouring part of a course on pedagogy (9). Recommendation 14 obtained a similar pattern of responses, instruction could be as a short course or as part of another course. Since this recommendation deals with devising scales and obtaining scores that differentiate at important decision points, it may be better to treat this topic in conjunction with interpreting and communicating assessment results, recommendation 15. Instruction in topics related to recommendation 5 may be better as part of instruction in pedagogy since subjective forms of assessment are integral to various curricular areas, and perhaps occur more in some subject areas than others. Although reliability is commonly thought of as a measurement concept, issues surrounding this in classroom assessment may require treatment within the curricular context (e.g., reliable assessment of written work as part of instruction in teaching language arts). Recommendation 16 (utility) deals with subjectivity also, and although not significant, did appear to favour instruction as part of pedagogy as well.

Reviewer responses to the nature of instruction for recommendations 14, 12, and 5 were similar, and frequencies did not differ significantly across the four approaches outlined. However, reviewers appeared to favour lecture and discussion, with frequencies of 10 to 11, over the other three approaches, which obtained frequencies of 7 or less. This suggests that issues of bias and prejudice in assessment, recommendation 12, could be dealt with in a lecture format, with support using some of the other approaches. Instruction related to recommendation 14 could also be addressed using lectures but with some hands-on support in laboratory work. Topics related to enhancing reliability for subjective forms of assessment, recommendation 5, could be taught by lecture and supported by the other approaches. As noted earlier, recommendation 16 deals with subjectivity as well, but responses here ranged only from 4 to 7, suggesting that all approaches are equally appropriate.

## Recommendations Rated Lower in Importance

Based on the reviewers' ratings of importance eight recommendations were interpreted as being less important than the nine discussed above (see Figures 4 and 7), although not all eight achieved ratings that were significantly lower. Since these recommendations were judged to be less important to classroom assessment they are treated more generally, and not all are discussed in terms of their implementation. There appear to be two groups of recommendations within these eight, the first five with means ranging from 3.4 to 3.2, which are above 3.0 on the scale which ranged from 1.0 to 4.0, and the remaining three with means of 3.0 to 2.6. The three recommendations with the lowest means reflect concerns for the psychometric properties of measurements as these are evinced in empirical and numerical procedures. These were clearly considered as less important to classroom practices, and are discussed separately and briefly below. The first five of the eight recommendations are:

> **Utility 16. Controlling the effects of subjectivity in observing and marking.** Both objective and subjective forms of assessment are necessary in classroom assessment, so teachers must be skilled in both. Teachers must be trained to use a number of different procedures for subjective evaluations, such as rating schemes, holistic scoring, and narrative descriptions. These procedures can be grounded in subject areas, where particular approaches may be prominent.
>
> **Utility 13. Making value interpretations of assessment results using referencing procedures.** The interpretation of scores, or other assessment information, is value-laden. Scores are interpreted in reference to norms, criteria, or the individual (self). Teachers must understand the issue of score referencing, and be able to apply the relevant techniques in specific situations. Teachers must also understand that part of the task is setting standards, which involves professional judgement.
>
> **Reliability 3. Increasing reliability through multiple quality assessments.** Students in schools must be given the opportunity to exhibit their

skills on several occasions, particularly if there is some doubt if the skill is mastered. Teachers must be able to obtain assessments of skills and knowledge in several different ways.

**Efficiency 17. Making maximum use of assessment time and effort.** The amount of assessment effort on the part of the teacher and of the students should be roughly proportional to the significance of the evaluation and the consequences of the decision. Teachers must know ways to speed up the assessment procedures for applied performance assessment including observation, selection-type tests (e.g., multiple choice), and longer constructed response assessments (e.g., written papers, research reports).

**Validity 8. Assessment for various purposes including instructional diagnostics.** Teachers must have the training to devise assessment procedures for a variety of purposes, such as for grading of students, individual and group instructional diagnosis, and evaluation of instruction. Teachers must also know when and how it is possible to accommodate more than one purpose, and to recognize when this leads to incompatible procedures.

Of these five recommendations, only recommendation 8 was rated significantly less important than any of the others (actually than recommendation 7 only). Thus, the ratings for the five recommendations were not much lower than those for the higher-rated nine, suggesting that they may be important enough to consider for implementation. This is particularly apparent for recommendations 16 and 3, which received very few reviewer ratings of not at all important (see Figure 4). Most of the recommendations received few reviewer responses of not at all important, and, as one reviewer commented, all the recommendations are important to teacher preparation but some are simply more important to deal with in detail. Recommendations 16, 13, 3, and 8 are related to, and were mentioned briefly in the discussions above of, recommendations 12 and 5, 15, 2, and 10, respectively. It may be possible to incorporate topics related to 16, 13, 3, and 8 in instruction related to the higher-rated recommendations. However, recommendation 17 is a distinct concept, and is the only one in the efficiency category: the ability of teachers to obtain the maximum information from the minimum of assessment time and effort. As one reviewer noted, "they will develop this soon enough when they are in the classroom". Others rated this recommendation as important, and it received the widest variation in reviewer ratings.

Recommendation 16 deals with controlling subjectivity in observing and marking student processes and products. It may be possible to combine recommendations 16 and 5 since 5 deals specifically with reliability of subjective assessments; it was also rated very similarly in importance (recommendation 12 is related, but refers to the broader issues of potential bias and prejudice in all types of assessment). Responses to the level of specificity of instruction for recommendation 16 were similar to those for recommendation 5: mildly favouring a theoretical basis (Figure 5). Reviewers were split as to whether instruction should be common or differentiated program for 16, whereas they were more supportive of common instruction for recommendation 5 (Figure 6). Responses to the method of delivery and the nature of instruction were not strongly distinguished for 16, but reviewers appeared to favour treating this as part of a course in pedagogy (pattern similar to that for recommendation 5).

Recommendation 13 deals with interpreting assessment results, which involves making value judgements and setting standards. Theoretical development was clearly favoured, with a mean of 3.3 which differed significantly from 2.5 ($p<.01$). There are many theoretical issues associated with this recommendation. But without specific classroom prescriptions it is probably destined to remain esoteric to teachers, and standards will be set the way they typically are now, implicitly and without the careful attention that they warrant, particularly at the higher grade levels. It is related to recommendation 15, which is concerned with more practical problems of communicating assessment results. Similar to recommendation 15, reviewers suggested instruction for recommendation 13 could be common for all students (although not as clearly so). There appeared to be mild support for delivery as part of pedagogy or part of one course and for lecture and discussion as the approach (responses were very similar to those for recommendation 15).

Recommendation 3 emphasizes the need for variations in settings and procedures to obtain assessment information. This is related to recommendation 2, but focuses on the range of the domain of generalization, and to making multiple, quality assessments. It is also related to recommendations 9 and 11, which refer to empirical validation of assessments from multiple and various sources. These latter recommendations emphasize the need for multiple pieces of assessment information, but instead to determine if they "hang together" within an assessment and across assessment procedures, essentially the validity notions of homogeneity and convergence. There were almost equal numbers of responses to recommendation 3 regarding theoretical basis versus specific prescriptions, and regarding common versus differentiated programs. Also, there were no clear differences in number of responses to the method of instruction or to the nature of the instruction.

Recommendation 17 deals with preparing students to be efficient in their assessment practices and to make maximum use of time and effort. Several reviewers thought this was important, although only limited time should be spent on it, whereas some thought it irrelevant. It was viewed more as a practical issue, and received mild support for specific prescriptions that could be imparted in a common or differentiated setting. Part of a course on pedagogy seemed to be slightly favoured, and any of the four approaches to instruction were acceptable.

Recommendation 8 is concerned with assessment purposes, such as diagnostics or grading, and is related to recommendation 10 (discussed above) in that different purposes may require different assessment structures, including scoring and interpretation (e.g., profile versus total test scores). Although considered of lesser importance, recommendation 8 achieved significance ($p<.05$) favouring theoretical instruction over specific prescription and also common programming over differentiation. This suggests that devising assessments appropriate to various purposes could be treated generally and as a common topic. Further, recommendation 8 achieved significance to the method of instruction ($p < .05$), reviewers appeared to favour both *part of one course* and *part of a course in pedagogy*, and to the nature of instruction ($p < .01$), reviewers favoured lecture and discussion.

The remaining recommendations in order of their rated importance are:

**Validity 9. Empirical checks of validity of the skills covered in classroom assessments.** It is important for classroom assessment that several tasks or items devised to assess the same skill or concept yield similar results with students. Teachers must know how to check for homogeneity. This involves identifying the assessment tasks or items which should "hang together" and checking student results to see if this is the case.

**Validity 11. Confirming the results of assessments by comparison with other assessments.** Previous assessments and other relevant evaluative information provide a check on assessment results, thereby giving external evidence for their validity. Teachers should understand the importance of external validation, and attempt to obtain several assessments of students' learning using different assessment procedures, and compare performances on these to see if there are glaring discrepancies.

**Reliability 4. Numerical procedures for estimating reliability in classroom assessments.** Teachers ought to confirm the adequacy of their measurements and check the reliability of some of the more critical assessments using straightforward numerical procedures.

Recommendations 9 and 11 refer to technical approaches for determining assessment the homogeneity and convergence of assessments, respectively, and, as earlier noted, are related to recommendations 2 and 3. Recommendations 9 and 11 were rated significantly less important than six of the seventeen recommendations, the highest rated six which include recommendation 2 but not 3. Recommendation 4 was rated as significantly less important than all recommendations except 8, 9, and 11, and with a mean of 2.6 was rated clearly less important than the rest; also only one reviewer rated it very important (see Figure 4). It recommends that numerical procedures be used to estimate the reliability of classroom assessments. This low rating is in keeping with the lower ratings awarded to empirical approaches to validation noted in recommendations 9 and 11. On the basis of this it is doubtful whether teacher preparation in assessment should include topics related to these three recommendations. It is clear from other research that empirical validation procedures are not part of the world of classroom teachers (e.g., Cole, 1987; Gullickson, 1986b; Stiggins & Bridgeford, 1985), and teachers simply do not use any statistical procedures to analyze their assessments (e.g., Gullickson & Ellwein, 1985; McLean, 1985; the case studies).

It may be possible to incorporate some of the notions underlying these three recommendations in the topics of instruction for other recommendations, but these should be restricted to the more practical aspects of increasing high-quality assessment and the amount of information when making decisions about students. The three recommendations are not discussed in detail, although for completeness, it should be noted that reviewers significantly favoured theoretical basis over specific prescriptions for recommendations 9 and 11, and common over differentiated instruction for all three. They also significantly favoured lecture and discussion for recommendation 4.

## Implications for Instruction in Classroom Assessment

Of the 17 recommendations identified for the model those rated as more important by the reviewers provided a basis for identifying the focus and content of instruction in classroom assessment for teachers in training.

### Focus of the Instruction

Nine recommendations were determined as being more important for teacher education than the others. The two rated as most important relate to producing assessments that are valid, assessments that support good instruction and that reflect the important learning outcomes. Clearly, instruction in classroom assessment must prepare teachers to evaluate students using procedures that reflect and support what they teach and emphasize, and that assess those learnings identified through curriculum guides and other documents as most important for their students. These learnings include the skills and knowledge related to subject matter content, as well as more complex thinking and reasoning skills. It is fairly straight forward to assess students on their spelling skills or their knowledge of number facts, but considerably less simple to assess students on their ability to interpret literary works or to apply scientific reasoning to phenomena around them. Most curricula outline learning objectives and identify approaches to teaching them, but it is less clear how teachers are to assess students considering the everyday requirements of the classroom. Teachers must be able to develop and use a variety of assessment techniques, including direct performance measures and paper-and-pencil tests, and these must correspond to and support the learnings expected of students. This is what Stiggins (1988a) means by the highest instructional priority. These assessment skills are clearly implied in the first two principles related to classroom assessments in the *Principles for Fair Student Assessment Practices* (*Principles for Fair Student Assessment Practices for Education in Canada*,1993), as well as referred to in the *Code of Fair Testing Practices in Education* (American Educational Research Association et al., 1988).

Rated next in importance was the recommendation that teachers can interpret properly the results of assessments, and can communicate both scores and interpretations effectively to students, parents, and others. Two recommendations rated somewhat lower in importance relate to this recommendation: making assessment results reflect accurately the purpose of the assessment, and obtaining the necessary scores. These should be incorporated in the instruction on this topic. The importance of the interpretation and communication of results is highlighted by the fact that the *Principles for Fair Student Assessment Practices* (1993) devoted some 17 principles to them. This includes the use of numerical summaries, but also various formats for grades and evaluative statements, to communicate student progress. Schools expect teachers to provide summary statements of student progress, usually via report cards, but the format for this varies widely (from narrative reporting at the primary level to letter grades and percentages at the secondary level, for example). Teachers must be skilled in translating assessment results into statements that clearly communicate what the information means. This must be done for subject matter learning and for social, affective, and other learnings that may not be bound to a particular subject area.

Several aspects of reliability were considered important for teacher preparation: that reliability be enhanced for assessments with major consequences for students, and that teachers know practical ways of improving reliability. The first of these requires that teachers understand how various assessments impact on decisions and on consequential evaluations, such as those embodied in grades. This is noted in the third principle related to classroom assessments in the *Principles for Fair Student Assessment Practices* (1993). Part of this aspect is identifying the purposes of an assessment and how it is to be used, and then developing the assessment procedures so that the results can reflect accurately the intended learnings (e.g., appropriate weighting of content). If the assessment results may have high-stakes consequences, reliability becomes more significant. An example where this may be problematic is when classroom assignments are included in the composite mark for a grade: these assignments usually are intended to be formative and generally are not very reliable. The overall reliability of results can be increased by including marks based on assessment procedures that are systematic and explicit, and by including more assessments of this nature--there are more opportunities for gathering information (this is also identified in the *Principles for Fair Student Assessment Practices*, 1993). Systematic procedures are more likely to be applied consistently across situations and students, and thereby increase the possibility of obtaining sufficient information for making decisions about students and communicating their progress. However, the procedures should be realistic for use in the classroom, and reasonable to expect of teachers.

Issues of personal bias and of controlling the effects of subjectivity in assessments were considered important but at a lower level. Reviewers also considered less important reliability and subjectivity in scoring student products and behaviours. Assessment procedures involving subjective scoring are necessary in the classroom because of the nature of many of the most important goals and objectives of the school program (e.g., writing ability, complex cognitive skills, affective outcomes). The issues associated with subjective assessment procedures may be best considered as part of the larger concern with controlling bias and prejudice in assessment. Bias in the content and the procedures of an assessment can result in skills other than those that are the object of the assessment affecting student performance. One possible bias is choosing a particular format for student responses, such as requiring sentence responses to arithmetic problems, while another is scorer bias (e.g., halo effect, logical error). All forms of bias may lead to unfairness in the use of the results. The effects on particular groups of students of the method of assessment (e.g., the type of response required of the student), the choice of language and the choice of elements of content, are not well known nor agreed upon by researchers (e.g., see Cole, 1981, and Jensen, 1980), but there appears to be evidence that these may affect groups of students differentially (e.g., Cole, 1981; Cole & Moss, 1989; *Educational Measurement: Issues and Practice*, Summer 1987). This is the basis for research on differential item functioning. Certainly there is evidence that wording and format in test questions affects student performance (e.g., Dawson-Sanders, Reshetar, & Shea, 1992; Wainer, 1989; White & Carcelli, 1982). The effects on students of teachers' personal biases and prejudices in the choice of assessment content and representations, such as ethnic and sexist stereotyping, are also difficult to determine, but should be controlled insofar as there are general guidelines for this. For example, it is practical in the classroom to attempt to make assessment materials and procedures give equal treatment to females and males, to students of different socioeconomic backgrounds and various racial and ethnic groups (this is noted in the *Principles for Fair Student Assessment Practices*, 1993). It is important that materials are either equally familiar to these groups, or that they are balanced so as not to disadvantage some students unfairly. It is also practical to make subjective scoring more consistent, such as by using the procedures suggested for holistic and analytic scoring of written work. Further, it is practical to design observation procedures that systematically cover all students under similar circumstances, and to reduce the incidental nature of much direct observation.

The remaining eight recommendations were rated lower in importance. One of these, controlling the effects of subjectivity in observing and marking, can be included as part of instruction in controlling for bias and making subjective assessments more reliable (discussed above). Reviewers considered least important the technical aspects of using numerical methods to check empirically the reliability and validity of assessments. This reflects the prevailing view that bringing anything but the most rudimentary numerical analysis procedures to assessment in the classroom is futile (e.g., Gullickson & Ellwein, 1985; McLean, 1985; Nitko, 1991a). Also considered of lesser importance were issues of score referencing procedures, and assessing for a variety of purposes. This may be due in part to the fact that grade levels 7 to 9 served as a basis for the model and the review: by grade 7 grading and reporting has become the primary purpose of assessments and most grading systems are defined in terms of scores accumulated from multiple assessments. Certainly by these grade levels, diagnosis and instructional review are at best secondary to the grading and reporting function.

## Approaches to Instruction

Preparing prospective teachers for classroom assessment can be done in many different ways. University-based professional schools was chosen as the context for identifying approaches to teacher preparation, and, within this context, four major features were outlined for the reviewers. University education is generally defined in terms of coursework that students must take to attain a degree or some form of certification. Coursework typically consists of lectures and class discussion, and may include laboratory or seminar components and practical experiences. This context posed some constraints on how instruction in classroom assessment was viewed: that it is part of a university program rather than a school system-based or apprenticeship program, for example. But implications based on the review are more likely to be practicable since, in the main, universities and colleges assume responsibility for initial teacher education in Canada and the US.

The instructional approach could vary for topics associated with each recommendation, but indications of the type of approach most appropriate to the recommendation gives some guidance. For example, for the two highest rated recommendations, some reviewers supported theoretical development of the topics whereas others supported specific prescriptions for classroom practice (although there were more who supported theoretical development). This suggests that instruction should provide some theoretical basis for devising assessments that support instruction and focus on important learning outcomes, but also it should give specific procedures for applications in the classroom. Instruction could introduce a model (or models) of learning and cognition that is grounded in research and that has been applied to the kinds of objectives prescribed by school curricula, and this model could then be applied to selected subject areas, such as language arts, mathematics, or science (models may come from many sources, but there is recent literature that gives guidance to this approach: e.g., Wittrock & Baker, 1991). Instruction would also include practical suggestions of how assessments would be conducted based on the model and on actual classroom material (e.g., Norris & Ennis, 1989, provides examples of how critical reasoning could be assessed in the classroom).

Reviewer support was generally split over theoretical development versus specific classroom prescriptions for eight of the nine higher-rated recommendations. Instruction on topics related to these recommendations would include theoretical development, but also practical classroom suggestions, although for six of these eight theoretical development appeared to be slightly favoured. For topics on practical ways to improve reliability, reviewers clearly supported classroom prescriptions over theoretical development. The amount of theoretical emphasis could vary for topics within each recommendation, but comments by reviewers indicated that topic development should be directly relevant to the work of teachers in the classroom.

It was informative to note that for all of the nine higher-rated recommendations more reviewers felt that instruction could be common for all education students rather than differentiated for particular subgroups. For two recommendations there was significantly more support for common instruction: communicating assessment results and making subjective assessments more reliable. This suggests that most topics related to these recommendations could be dealt with in coursework available to all prospective teachers. However, there was also substantial support for coursework differentiated according to the program level, such as elementary and secondary, particularly for topics on assessing higher-level cognitive learning and practical ways to improve reliability. It is reasonable to differentiate instruction for program levels and subject specialties regarding what is important to assess and how this relates to the development of the child, but probably less so for practically enhancing reliability. This depends in part on how instruction in classroom assessment is conducted: if it is a separate course specifically devoted to assessment many of the topics regarding validity and reliability could be generic, but some of the instruction in assessment would then be necessary to develop in other courses, such as those on pedagogy and curriculum. For example, it is possible to teach generic approaches to the assessment of higher-level thinking skills, but it is likely to be more effective if these are further developed within the context of instruction in language teaching and science teaching, and in the other curricular areas. Finally, teachers require the skills to prepare their own assessment materials, which must be appropriate to the students they will be teaching, so they must understand what is important to assess at these levels, and how best to assess these learnings. The reviewers supported this view, and suggested that instruction as part of a separate course and as part of a course in pedagogy is appropriate for these aspects of validity.

In general, reviewers did not support the use of brief or discrete instructional modules for any of the recommendations. Rather, for the higher-rated recommendations, they generally supported making instruction part of a course on assessment and part of instruction in pedagogy, with perhaps slightly greater support for instruction as part of pedagogy. This appears to be based on the need to provide instruction that is more protracted than a one-shot presentation and that it should be integrated with and supported by instruction in the whats and hows of teaching. They also generally endorsed the lecture and class discussion as the nature of instruction, which is in keeping with their view of the importance for providing the theoretical basis for topics in assessment. However, there was also considerable support for other forms of practice and experience. This indicates that prospective teachers should have ample opportunity to obtain practice in preparing and conducting assessments, and this practice would include laboratory-type work as well as direct experience in classroom settings. There are large practical problems with providing in-school experience on assessment since it is difficult to provide settings that correspond directly and systematically with all aspects of instruction in assessment. It may be possible to provide simulated and laboratory settings for some topics, and reserve classroom experiences for those topics for which this is most necessary. This is difficult to determine, but it would appear reasonable to have students design and prepare assessment procedures and materials in a laboratory setting, and to try some of these with school students--perhaps selecting some paper-and-pencil techniques and some that require direct observation. It may be possible to integrate these types of school experiences, with the enhancement of assessment as the goal, with other school experiences that are part of teacher preparation.

# VI. GUIDELINES FOR TEACHER PREPARATION IN CLASSROOM ASSESSMENT

The purpose of the study was to develop an instructional component for the preparation of teachers in classroom assessment. This component was to reflect measurement and evaluation principles, yet be realistically applicable to the classroom. It was to set forth the focus, structure, and design for delivery of a preparatory program for education students in assessment that was both practicable and theoretically defensible.

There were two phases to the study. The first was to obtain an indication of what was realistically possible in the classroom in terms of assessment. The assessment practices of four exemplary teachers at the junior high school level were determined by observing and interviewing them, and analyzing their assessment materials. These case studies are described in Chapter III. The second phase of the study involved identifying the theoretical basis for assessment, and forming recommendations for teacher preparation in classroom assessment that were structured on assessment principles. The recommendations also accommodated the reality of the classroom context as determined from the case studies. This is presented in Chapter IV, and summarized in Appendix D. The second phase was to further establish the appropriateness of the recommendations. They were submitted to a critical review by a number of educators ranging from measurement specialists to classroom teachers. This review is reported in Chapter V.

Chapter VI provides a synopsis of the findings from the case studies, including a number of implications these have for preparing teachers in assessment. This is followed by a brief summary of the basis for the recommendations given in Chapter IV, and conclusions based on the review of the recommendations in Chapter V. Chapter VI concludes with a set of principles to guide the development of the assessment component in teacher education, and a description of what this could entail including how it would fit into the overall teacher preparation program.

## General Findings and Implications From the Case Studies

There were six underlying themes that emerged from the case studies. They were expressed in broad competencies that teachers require if they are to carry out effective assessment in the classroom. In the discussion that follows, the limitations of the implications are noted first. This is followed by a summary of the six competencies, and then by a more detailed discussion including the rationale on which the competencies were based and suggestions for topics of instruction for prospective teachers. This discussion in under the headings of Using assessment information effectively, and Designing and conducting appropriate assessments.

### Limitations of the Case Studies

Four experienced and exemplary teachers of science and social studies at the junior high school level were selected for the first phase of the research, the case studies. The implications for the preparation of teachers in assessment were based on observations of these four teachers in the classroom, on interviews with them, and on appraisals of their assessment materials (see Chapter III). The review of literature provided further grounding for the implications, and suggested that they can be applied more generally to teachers at other grade levels and of other subject areas. However, there are clear limitations, based on the fact that the number of teachers was small, and these teachers were not representative of teachers generally. Further, differences could be expected between teachers of various subject areas, and of different grade levels. The number of teachers in the case studies was too small to determine what these differences are.

## Teacher Competencies in Classroom Assessment

The first three competencies deal with the purposes of assessment and the use of assessments to derive statements of student progress. The latter three competencies deal with the development, administration, and scoring of assessments. In summary form, the six competencies are:

1. Teachers must be able to identify how assessments are to be used, what purpose or purposes they are to serve, and how this impacts on the design of the assessments.

2. Teachers must be able to make fair and defensible summative statements of student progress. Teachers must be able to form composite scores that maintain the importance of the components, and that can be translated, with a minimum of ambiguity, into qualitative summaries based on standards of performance.

3. Teachers must be able to identify aspects of learning and development on which it is important to report student progress, and to communicate this effectively to students, parents, and others. Teachers must be able to provide information that assists students in their development .

4. Teachers must be able to design and conduct assessments that are appropriate to the goals and objectives of the school program and to their instruction, and that therefore are legitimate for the anticipated purposes.

5. Teachers must be able to develop and write questions and test items that validly reflect the learning outcomes, and that are clear to students and free from potential biases and mechanical errors.

6. Teachers must be able to conduct systematic and fair observations of students that focus on the learning outcomes, particularly if this information is to be used for consequential assessments.

## Using Assessment Information Effectively

There was considerable variation among the teachers in the case studies as to what importance they attached to various assessment purposes, but in general they identified assigning grades as the most important, followed by evaluating instruction. Diagnosing individual student and group needs received some support. These findings are in keeping with those of others (e.g., Webster, 1987; R. J. Wilson, 1989). Variations across grade levels could be expected, with teachers at the lower grades placing less emphasis on grading and reporting and more on diagnosis, whereas those at the higher grades, including junior high, placing more on grading and reporting.

Assessments often serve multiple purposes, and there is little doubt that, besides forming part of a composite grade, teachers use assessment results to inform their instruction and to identify areas of students' strengths and weaknesses, and also to guide and focus student learning and effort (the case studies; Brookhart, 1992). Assessments can be designed to serve the purpose of providing an overall indication of student achievement to be used as part of a composite score for summative evaluation, but assessments that do this well may not be as effective for deciding where instructional effort should be placed. Conversely, not all assessments provide information that is appropriate for including in a composite grade: for example, some are too detailed and may inappropriately emphasize selected material, others may be embedded in students' learning and indicate more interim skills and errors rather than overall achievement or performance. Good pedagogy suggests that students should feel free to experiment, and should be encouraged to do so, without fear of being graded on these attempts which may well be off topic or even incorrect.

This is an issue of validity, the validity of assessment results in informing a decision. In the document *Principles for Fair Student Assessment Practices* (1993) it is discussed in terms of summative evaluations and grades under the heading Summarizing and Interpreting Results. Here, it is probably not stated strongly enough since, for example, some teachers accumulate as part of grades assessments of almost everything students produce including that which is clearly expected to be formative. Principles 4 and 5 under this heading do address combining information from multiple sources: the problems associated with obtaining a balance of the learning outcomes, and in combining disparate skills and other characteristics such as effort in forming a score. These problems are not simple, and range from philosophical issues (e.g., Frisbie & Waltman, 1992; Terwilliger, 1989) to technical and statistical ones (Oosterhof, 1987; Thayer, 1991). However, there are further problems encountered in classroom assessment, such as how to deal with assignments and other forms of student work which are intended as learning activities, but represent accomplishment on the part of students. There are affective characteristics, such as effort, care and neatness, attitude to subject and to school, cooperation, and initiative that may come into play in the grades awarded students. These characteristics may be implicitly considered in the total mark for an assignment (e.g., by giving marks for neatness or presentation or if a particular format is used, or taking off marks if the assignment is late), or more explicitly where teachers use this type of information in deciding on the grade for a student whose composite score is on the borderline between two grades.

Teachers part company with measurement experts on what should be included in forming composites for grades--they typically include data from sources other than achievement measures (Friedman & Manley, 1991; Stiggins, Frisbie, & Griswold, 1989; Waltman & Frisbie, 1993). Characteristics other than subject matter understanding are clearly important, and this is recognized by educational systems in curriculum and policy documents, by teachers and schools in report cards (e.g., they often include separate categories for reporting effort, Friedman & Frisbie, 1993), and by parents in their expectations of schools (e.g., Waltman & Frisbie, 1993). But what the appropriate ways are of assessing and communicating these kinds of learning's is not as clear, and certainly the subject of debate. Apparently most teachers view grades as that which students *earn* for the work they do, "grades functioned as the coin of the realm" (Brookhart, 1992, p. 4), rather than more narrowly as a measure of academic achievement in a subject area. Arguments for the inclusion of effort and other factors in marking students' work and in forming grades are based on the consequential aspect of grades. Grades are used in a variety of ways, not simply as measures of students' academic ability, and have implications for students' views of themselves and their views of the school environment (e.g., what is important and what they are to do in the classroom, Brookhart, 1993; Pilcher-Carlton & Oosterhof, 1993). Effort is clearly something that is valued and therefore rewarded.

These issues are fundamental to instruction of prospective teachers. Grading and reporting is the major function of assessment in the classroom, and teachers need guidance in how they are to perform this function. But guidance must go beyond categorical statements that grades must be based on measures of cognitive skills and communicate only understanding of the subject content, as is currently the case in most measurement textbooks (e.g., Gronlund & Linn, 1990; Ebel & Frisbie, 1991). There are procedures for assessing and communicating affective learning (e.g., L. Anderson, 1981) and these assessments can be reported separately from cognitive academic learning, as is done on some report cards (Friedman & Frisbie, 1993). However, such characteristics as effort and care are difficult to separate from achievement particularly when performance is measured by actual behaviours and products. Teachers, and students also, believe that fairness is key to grading, and the quality of student work is the primary criterion for this fairness, thus often confounding effort and ability since the evaluation of quality often includes the care and effort with which the student prepares the product. Further, students may produce good quality products in the sense that they are thorough, complete, and well presented, but not necessarily understand many of the concepts that underlie the product. For example, as part of a science topic on heat transfer and convection, one of the teachers asked students to describe the heating system in their homes; included in the report were to be such things as type and location of the furnace and diagrams of air flow. This task allowed students to exercise a number of science process skills, such as identifying the scientific principles that are operating in a practical setting and communicating in a variety of formats. A student could provide a fairly complete description of the home heating system, yet not understand the principle relating temperature to air density, the effect this has on air movement, and how technology is designed to accommodate the effect.

The difficulties surrounding the meaning of grades are not readily resolved, but as Griswold and Griswold (1992) state, "teachers need to be party to redefining the notion of achieving success in the classroom as well as to measuring it" (p. 4). Grades are not used only as a measure of academic performance for predicting future school performance; they serve as rewards, and as indicators of diligence and effort. This must be addressed in preservice teacher education: the issues noted above must be made clear, and the implications of particular assessment procedures must be understood. We should stop short of exhorting future teachers to base grades solely on measures of academic achievement as this seems to consistently contravene the practices of even the best educators in the field. Guidance should be given teachers on procedures to obtain composite scores and form grades that are systematic and equitable, and that can be explained to students, parents, and others so that it is clear what is meant by them. For example, if aspects such as student effort are to be included in grades this should be clearly stated.

Teachers in the case studies were conscientious in marking assignments as well as tests and quizzes (most were marked by the teacher), and returning these to students promptly (most within one or two school days). Feedback was provided to students in the form of the mark awarded, and also written comments and suggestions. Two of the teachers provided the students with class averages and other information on some of the tests and assignments. All this information can be useful to students, but some of it is intended to be formative, such as the comments and suggestions. Some of the information was to be used in a summative sense as part of the overall indication of performance, such as the class averages.

Schools tend to have fixed progress reporting periods, usually three or four times annually, but the format and content of these varies widely (the cases studies; Friedman & Frisbie, 1993; Schulz, 1993; Stiggins, Frisbie, & Griswold, 1989). One school in the case studies provided considerable detail, including on aspects such as homework, daily assignments, tests, projects, effort, participation, and behaviour. This school also required teachers to provide a breakdown of how composite marks were obtained. There are several levels of reporting, of which summative grades is one; others range from marks and comments on assignments to summary statements of strengths and weaknesses or problems areas. It is difficult to determine what should be contained as part of a term report or report card, and suggestions vary from detailed lists of skills attained to global indicators of performance in subject areas. There is evidence that parents want summative statements, and find too much detail confusing and uninformative. They also want indications of problem areas and of particular strengths and skills. The broad principle that guides the form and content of reporting is noted in the *Principles for Fair Student Assessment Practices* (1993): communication must be understandable and encourage learning.

Three competencies were identified to capture generally what teachers should be able to do in using assessments effectively for the purposes for which they were intended. Some instructional topics that are relevant and that should be addressed in teacher preparation are noted after each competency.

1. Teachers must be able to identify how assessments are to be used, what purpose or purposes they are to serve, and how this impacts on the design of the assessments. In particular, they must be able to identify assessment designs that are appropriate for grading and summative reporting purposes, and those that are appropriate for diagnosing learning difficulties, modifying instruction, and other formative purposes.

Teacher preparation must include instruction on how to distinguish among different assessment purposes, particularly between summative and diagnostic purposes, and to identify characteristics of assessment appropriate to each. For example, if an assessment is to be used for summative evaluation, such as part of a grade, the content topics and skill levels must be balanced according to the objectives of the instruction. Tables of specifications or test blueprints are a technique which can be used to do this, although they are not popular with teachers. The overall score is the focus of interpretation. If an assessment is to provide diagnostic information it must be designed to give information on a number of skills that can be identified as fundamental to broader or superordinate abilities. These have been outlined in certain subject areas, such as reading and mathematics, but task analysis is one procedure that can be used to identify subordinate skills where these are not available (science and social studies, for example). Assessment information on one skill must be sufficiently stable and differentiable from other skills to warrant interpretation and possible instructional modification. Subtest scores become the basis of interpretation, and must be reasonably reliable and valid.

2. Teachers must be able to make fair and defensible summative statements of student progress. Usually these are in the form of qualitative written statements or grades, and require that teachers determine standards of performance. Teachers must be able to form composite scores using procedures that maintain the intended importance of the components, and that yield scores that can be translated into qualitative summaries with a minimum of ambiguity.

Teachers must be instructed in how to form composites scores using procedures that maintain the relative importance of each component score by applying appropriate weights. The procedures must take into account the spread of scores for each component. Measures of spread are not likely to be used in the formula for the composite, nor are they necessarily effective (Thayer, 1992), so teachers must develop scoring procedures and award marks so that ranges in scores are obtained that are suitable to differences in performance and that allow forming composites that do not obscure the differences in scores on each component.

Teachers must be taught procedures to define performance standards that make them explicit and communicable, and that can be used to interpret student performance to the students and to others. This involves setting criteria for performance, and defining procedures to obtain defensible and fair summary qualitative statements, including grades, checklists, and narrative accounts; communicating the summary statements and how they are to interpreted by indicating clearly what was being assessed, how scores and grades were obtained, and what and how criteria were applied.

3. Teachers must be able to identify aspects of learning and development on which it is important to report student progress, and to communicate this effectively to students, parents, and others. Reporting consists of more than summary statements of student progress. It consists of informing others clearly on how well a student is doing, which often includes a grade and qualitative statement, but varies according to the needs of the student and the audience. Teachers must be able to provide information that assists students in their development .

Teachers must be instructed in what are considered the most important learning's for students at certain grade levels, and in how to state these clearly and in ways that they can be understood by students in particular, but also by parents and others. They must know how to devise scoring procedures for student behaviours and products which identify what is to be achieved, and that limit the confounding of affective and other factors with measures of knowledge and understanding. The factors that are considered in the scoring should be consonant with the goals of instruction, and should be delineated clearly and communicated to students, but also should be available readily to others such as parents and school administrators.

Certain aspects of overall learning can be summarized in a grade or some other similar qualitative indicator. Other aspects may require different formats, and supporting commentary beyond a single indicator. Teachers must be instructed in using all formats appropriate to the grade level, area of learning or development, and needs of the child.

## Designing and Conducting Appropriate Assessments

The methods used by the four teachers to measure achievement emphasized paper and pencil tests. This was followed in emphasis by regular homework assignments (mostly paper and pencil), and performance assessments (including labs, projects, etc.). This is in keeping with what was reported in other studies (e.g., Bateson, 1990; Gullickson, 1985; R. J. Wilson, 1989), although the teachers in the case studies appeared to place slightly greater importance on procedures other than paper and pencil, such as performance assessment and group assessment methods. This may have been a function of the subject areas, science and social studies, but may also have been because the teachers selected were exemplary and thus more likely to use a wider range of assessment tools. Classroom assessment practices are related to grade level, with observations, assignments, and work samples used more at the lower grades, and assignments and formal tests used more at the higher grades (e.g., J. O. Anderson & Bachor, 1993). Thus, it is expected that teachers at the senior high level would use even more paper and pencil tests than teachers at the junior high level (than those in the case studies), and teachers at the elementary and primary levels would use some paper and pencil tests, but would use more informal assessments, those based on work samples and on direct observation.

The overriding concern with teacher-made assessments is that they greatly overemphasize the simpler forms of cognitive understanding, such as repeating stated facts and principles. This has been reported for classroom tests by researchers such as

Fleming and Chambers (1983), Gullickson (1985), Haertel (1986), and Stiggins, Griswold, and Wikelund (1989), and was confirmed in the case studies. The effect of this is to misrepresent the goals and objectives of the school program, and to misdirect the focus of students' effort since students readily identify what is important based on what is being assessed (Crooks, 1988).

The methods used to measure affect were not as definitive since two of the teachers in the case studies reported that they did very little of this, and questioned whether it was appropriate to assess affect and report on it separately from subject matter achievement. Two of the teachers reported that they assessed affective characteristics, and used both brief questionnaires and observations. They used questions like "What do you like about science (social studies)?" and "How important is science to society?" Although teachers readily acknowledge the importance of the affective domain in education they are uncomfortable with assessing it generally as well as in subject areas (e.g., Webster, 1987). The problems associated with incorporating affective aspects along with academic achievement in an assessment are noted above. However, program and curriculum documents define affective learning's for students, and school systems expect teachers to assess and report student progress in this area (e.g., Schulz, 1993).

The primary criteria teachers in the case studies used for selection of assessment methods were that they fit the purpose of the assessment and match the intended outcomes of instruction. Since the expected outcomes for our students include not only subject matter knowledge, but a variety of academic skills, attitudes, and behaviours, assessment procedures applicable to these types of learning's should be covered in a teacher preparation program.

Teachers in the case studies produced paper and pencil tests that were generally free from mechanical errors and from gender and ethnic/racial bias. Some of the directions to students were not clear, and the marks to be awarded and how written response items were to be marked was often not indicated. There were some mechanical problems with matching and true-false items. The directions for some matching exercises were not clear, and the basis for matching was not typically stated, although it could be inferred from the lists. More problematic was the inclusion in a list of elements that were not related: for example, a list might contain the names of geographical features and cultural artifacts, or some elements may be names of things and others labels for processes. This allowed students who could distinguish terms that represent entities from those that apply to actions and make correct matches, or select responses from a subset of those given in the second list, without knowing the precise meanings of the terms. As noted above, there was a paucity of items that required anything more of students than direct recall. Context dependent items are argued to be well suited to assessing more than recall or direct extraction of information (e.g., Haladyna, 1992), but teachers find these difficult to develop. Thus, teachers require considerable training and experience in writing test items in a variety of selection and supply formats.

Teachers rely heavily on direct observation for their assessment information, particularly at the lower grades, but also for assessing affective outcomes at all levels. Although some skills may be better assessed using tests, there are many which require observations of student behaviours: for example, process skills in science such as measuring and affective behaviours such as cooperation. Evidence from Stiggins (1986a) and from the case studies, though, indicated that teachers do not tend to use well-defined and systematic procedures for these observations. Oral forms of assessment are infrequently used except as part of ongoing classroom instruction, although teachers in the case studies indicated conducting oral assessments for some students who were less able to handle written tests. Oral questioning during instruction was typically not systematic or structured, but the teachers were aware of whom they questioning and they attempted to include all students. Performance assessment is an important part of classroom assessment at all grade levels and in most subject areas. This includes assessment of both student behaviours and their products, although at the higher grade levels the product is usually the focus of assessment and skills are inferred from this. An example from the case studies was using the assessment of responses to laboratory exercises and of laboratory reports to infer students' science process skills. Student products varied in the case studies, and they included not only written research reports but also oral and video reports, posters, and other forms of display. The teachers frequently devised and used a scale or multiple scales to assess these products.

Teachers use group work frequently in the classroom. This occurred in the case studies for most science laboratory assignments and for some papers and projects in social studies. Thus, this instructional format should be accommodated in classroom assessments. There are problems with assessments based on students working in groups, such as assessing groups outcomes and using this to grade individual students. It is possible to structure both the group situation and the assessment procedures so that the assessment is more likely to correspond to the intended learning's and is more equitable. An example of this is in providing structured, usually simulated, settings involving science problems (experiments, as these are often defined in the classroom), and systematically observing the behaviour and performance of individuals in the group.

The teachers reported their own experience as the most important source for assessment information, even though they all had attended many inservice sessions and taken university courses while they were teaching. They also made relatively little use of assessment materials from outside sources, including commercial tests and items which accompany textbooks. This suggests that preservice assessment training is all the more important since it is difficult to change classroom practice once teachers have begun teaching. The teachers reported allocating most of their assessment time to selecting and developing their own assessments and scoring assessments, and on administering assessments and providing feedback. This shows that teachers spend little time on evaluating their assessment procedures, and certainly this puts into question the utility of instructing teachers in the use of complex numerical analyses such as item analysis.

Three competencies were identified to capture what teachers should be able to do in translating the goals and objectives of the school program into procedures for assessing students fairly and systematically. Some instructional topics that are relevant are noted after each competency.

**4.** Teachers must be able to design and conduct assessments that are appropriate to the goals and objectives of the school program and to their instruction, and that therefore are legitimate for the anticipated purposes. These assessments must represent the full range of knowledge, skills, attitudes, and behaviours expected of our students both relative to subject areas and across the curriculum. Teachers must be able to identify clearly the learning outcomes of interest, and procedures that are valid for the assessment of them.

Teachers at all grade levels must be instructed in how to design procedures for the assessment and communication of aspects of student learning identified as important for our students. Subject matter knowledge and understanding is the focus of much of the school program, and serves as the basis for instructional outcomes particularly at the secondary school level. But here, and more so at the lower grades, there is also much of the intended learning that cuts across curriculum areas, such as skills in oral and written communication. The learning outcomes include processes associated with, or emphasized in, subject areas: in science and social studies these are skills such as observing, recording, organizing, and communicating information; forming and testing hypotheses; controlling variables and experimenting; drawing conclusions from data; and forming implications. This also includes skills that are reflected in most curricular areas, such as oral and written communication, other forms of expression and communication (e.g., pictures, diagrams, graphs, demonstrations), and critical thinking using principles of logic and reasoning. Teachers must have ample instruction in how these outcomes can be identified clearly, and how procedures can be structured to validly assess these outcomes.

Teachers must be prepared in how to design procedures for the assessment and communication of affective and behavioural aspects of student progress. Some of these would be associated with subject areas, although most would be appropriate more generally: for example, in science and social studies this includes open-mindedness and withholding judgement, questioning assertions and seeking information, being critical of one's own beliefs. This also includes characteristics of a generic nature, appropriate to all program areas, such as effort, motivation, honesty, integrity, care and neatness, cooperativeness, respect for and acceptance of others and others' beliefs and customs, tolerance of differences, initiative, personal and social responsibility, and creativity and ingenuity. Procedures to assess these types of learning and development are not as developed as those of a cognitive nature, but teachers must have instruction in translating some of these to observable and assessable behaviours and outcomes.

**5.** Teachers must be able to develop and write questions and test items that validly reflect the learning outcomes, and that are clear to students and free from potential biases and mechanical flaws.

Teachers must be skilled in the specifics of preparing questions and paper and pencil test items of various formats. Teachers must have instruction and experience in the development of questions that "get at" the desired skill, that are clear and understandable to students, that are free from potential biases and prejudices (e.g., gender, ethnic origin, social group, geographic region), and that are free from mechanical flaws. In general, the question formats include selection type (e.g., true-false and other alternate response, matching, multiple choice) and supply type (e.g., completion and short answer, restricted response essays). These are well described in most measurement textbooks, but teachers should have ample experience in using the guidelines set forth for them.

Teachers must be instructed in preparing complex test items and context-dependent exercises of various formats, items that require more than extraction of information. The item material should include various forms (e.g., textual, graphic, numeric, pictorial) and the questions should also require varied responses. Teachers should also have ample experience with procedures for marking students' written work (and work in other forms) since this is such a large part of our school program.

**6.** Teachers must be able to conduct systematic and fair observations of students that focus on the learning outcomes, particularly if this information is to be used for consequential assessments.

Teachers must be instructed in identifying clearly the behaviours expected of students as indicators of certain skills and attitudes. For example, as evidence of open-mindedness, it may be necessary to identify a commonly held misconception (say, in science, such as heavier objects fall faster than lighter objects), confronting this with contradictory evidence (have students conduct a test of this), and observing change in student behaviour (prediction of rate of fall of objects of different weights).

Teachers must also be skilled in preparing and using procedures for the systematic observation of students. This includes preparing common or equitable settings in which students are to exhibit various skills and behaviours. These can be incorporated as part of performance assessments, but must provide for observation of all students under reasonably similar conditions. A variety of techniques can be presented for conducting efficient and systematic observations, such as time sampling, random selection of students for daily observations, or preparing a number of assessment stations and having students circulate from one to another while being observed.

Teachers should have some experience in adapting assessment procedures for use with student groups. Teachers make considerable use of student groups for both instruction and assessment. Further, observing students working in small groups is important for assessing certain affective characteristics, such as cooperativeness, responsibility, and having respect for the ideas of others, but students must have equitable situations for displaying these behaviours. An example of this is when students are to designate roles and tasks within the group: it is much easier to observe responsibility on the part of the chairperson than on other members of the group. Groups are also important for teaching and assessing decision making and problem solving, but individuals must be assessed to determine if important skills are developed. This may be possible to do using observation, but again it would be necessary to carefully devise procedures that allow each student to exhibit the important learning's.

## Specification and Review of the Recommendations

The characteristics of good classroom assessment practices were identified under four headings: reliability, validity, utility, and efficiency. These headings formed the theoretical structure to organize the recommendations for teacher preparation in assessment. The basis for the 17 recommendations, as well as detailed statements of them are presented in Chapter IV (summary statements for the recommendations are given in Figure 3, Chapter V). The recommendations are also given in the summary below.

A document was prepared that specified each recommendation in some detail and provided the rationale for it. This document also indicated the procedure to be used for systematically reviewing the recommendations as to their importance and how they might be implemented (this is duplicated in Appendix D). The document was presented to 16 educators for review: four measurement specialists (three university professors and one at the division level); five curriculum specialists (two in social studies and three in science, three university professors and two at the division level); one principal and one division administrator; and five experienced teachers (four of whom participated in the case studies). The focus of the review was on the importance of each recommendation for teacher preparation, and how topics related to the recommendation might be implemented.

The review is discussed in Chapter V, and details of the analysis are given in Appendix D. The reviewers rated the importance of each recommendation (Figure 4), and these ratings were used to order and group the recommendations for discussion. The reviewers further responded to four features of program delivery: level of specificity of instruction (Figure 5); common instruction for all education students or differentiated for some groups (Figure 6); method of delivery of the instruction (e.g., part of one course, part of instruction in pedagogy); and nature of the instruction (e.g., lecture, laboratory, in-school experience).

The four characteristics of good classroom assessment used to organize the recommendations do not form a taxonomy of classification, but served more as convenient and well-recognized categories that highlighted important considerations in measurement and evaluation. Validation is clearly at the heart of assessment, and seven recommendations were identified in this regard. Five recommendations pertained to reliability, four to utility, and one to efficiency. For discussion the utility and efficiency recommendations were combined. The basis for the recommendations and the results of the review are summarized below in the order reliability, validity, and utility and efficiency.

## Limitations of the Review of the Recommendations

The number of reviewers was sufficient to provide a basis for general comments regarding an approach to instruction related to the recommendations. However, there may be substantial differences in what different educators perceive to be important for teacher preparation in classroom assessment. For example, curriculum specialists may view assessment needs differently from measurement specialists, and both groups may differ in their views form those of practicing teachers. This was not pursued as general indications were sought, and the number of individuals selected from each educator group was too small to permit meaningful comparisons among them.

## Reliability

A criterion-referenced model was considered most applicable for classroom assessment purposes, but procedures based on norm-referenced theory could also be used. Classroom assessments are not usually clearly one or the other (e.g., Brookhart, 1992; Terwilliger, 1989). In general it was argued that technical procedures for reliability estimation were not appropriate, and, in any case, could not reasonably be expected to be used by teachers, nor was it clear what teachers would do with the results. The five recommendations reflected this reality, and emphasized the practical side of enhancing reliability, rather than its measurement. The reliability recommendations are:

1. **The importance of reliability for high-stakes assessments.** Teachers must be able to identify the main purpose of an assessment and its consequences for the student, and enhance reliability accordingly.

2. **Practical ways of improving reliability in the classroom.** Teachers must know how to enhance reliability for particular classroom assessment practices by making the procedures more explicit and systematic, and by increasing the amount of high-quality information gathered.

3. **Increasing reliability through multiple, quality assessments.** Teachers must be able to obtain assessments of skills and knowledge in several different ways.

4. **Numerical procedures for estimating reliability in classroom assessments.** Teachers ought to be able to check the reliability of some of the more critical assessments using straightforward numerical procedures.

5. **Reliability of subjective forms of assessment.** Teachers must understand that subjective forms of assessment are prone to unreliability and to personal bias and know practical procedures which can ameliorate these problems.

Recommendations 1 and 2 were rated by the reviewers as being more important than three of the others, and recommendation 5 was rated as more important than one other. Recommendation 3 was rated as relatively unimportant, and 4 as the least important of all 17 recommendations. The importance of 1 reflected the present awareness of high-stakes assessment, and 2 is a practical recommendation on how to improve reliability in the classroom. Recommendation 5 is simply an elaboration on 2, since it refers to one area of assessment where reliability can be enhanced in specific ways. Recommendation 3 is probably better considered as an issue of validity, except insofar as it supports the notion in 2 that reliability is enhanced by increasing the amount of high-quality (systematic, consistent) assessment procedures, not necessarily by more assessment. The low rating of recommendation 4 is not surprising since numerical procedures are not well received in classroom assessment generally.

Recommendations 1 and 2 should clearly be part of teacher preparation, and 5 likely also ought to be included. Inclusion of recommendations 3 and 4 is questionable on the basis of the ratings. These recommendations entail topics which Schafer (1991) identifies as important, such as understanding error of measurement and using reliability to evaluate the quality of assessments, suggesting that numerical estimation procedures are appropriate. On the other hand, Stiggins (1991a) argued that the numerical and other technical procedures are not essential to classroom assessment and teacher preparation, when these topics are compared with other characteristics and requirements, given the importance of other topics and the limited time that can be devoted to assessment in teacher preparation. Airaisian (1991), too, emphasized the need for practical orientation in teacher training. However, there is argument supporting the need for teachers to understand the notions of instability in measurement, and how this can be estimated, particularly in high-stakes uses of scores.

The procedures and approaches that were suggested as most appropriate for development of the topics identified in recommendations 1 and 2 (and to a lesser extent, 5) are that there should be some theoretical development but this should be made specific to the classroom. Specific prescriptions are particularly important for topics related to recommendation 2. Much of this could be taught to education students in common, with some differentiation through illustrations and examples for elementary versus secondary program students. It was felt that much of this could be part of a course designed for assessment, but that there should be support and further development as part of preparation in pedagogy.

## Validity

The most important characteristic of assessment is clearly validity. This includes substantial, structural, and empirical components of assessment results as well as personal and social consequences in the use of an assessment. The older categories of content, criterion, and construct validity are more for convenience than meaningful breakdowns of validation approaches (e.g., American Educational Research Association et al., 1985; Angoff, 1988; Messick, 1989b). The seven recommendations under this heading are based on the more modern view, but do not reflect the components directly. The notions of structural and empirical validation are blurred, and often coincide in practical classroom applications. The validity recommendations are:

Recommendations 6 and 7 were rated by the reviewers as most important of all 17 recommendations. This is in keeping with recent views that the essence of teacher preparation is in assessing the most important learning outcomes and in ways that inform and reinforce their further learning (e.g., Nickerson, 1989a; Shepard, 1989; Stiggins, 1988a, 1991a; Wiggins, 1989a, b; Wolf, Bixby, Glenn, & Gardner, 1991; Yager, 1989).

These recommendations reflect primarily the substantive aspect of validity, but also highlight the consequential nature of assessments.

6. **Increasing assessment of important, higher level cognitive learnings.** Teachers must have the knowledge and skills to design and conduct classroom assessments that are most appropriate to the most important learning outcomes, including processes of higher-level thinking and reasoning.

7. **Assessment that supports and informs good instruction.** Teachers must be able to use assessment procedures that support the most important learning in classrooms, those that should be emphasized in class and on the assessments, including paper and pencil techniques, but more importantly, systematic observation and applied performance assessment.

8. **Assessment for various purposes including instructional diagnostics.** Teachers must be able to devise assessment procedures for a variety of purposes, such as for grading of students, individual and group instructional diagnosis, and evaluation of instruction. Teachers must also know when and how it is possible to accommodate more than one purpose, and to recognize when this leads to incompatible procedures.

9. **Empirical checks of validity of the skills covered in classroom assessments.** Several tasks or items devised to assess the same skill or concept should yield similar results with students, and teachers must know how to check for this homogeneity. This involves identifying the assessment tasks or items which should "hang together" and checking student results to see if this is the case.

10. **Making assessment results reflect the focus and purpose of the assessment.** Teachers must be able to obtain scores from assessments that correspond to the nature and purpose of the assessment. This involves identifying in advance the approximate importance, and hence the weighting, of various areas of content and levels of skills (essentially producing a table of specifications for the assessment).

11. **Confirming the results of assessments by comparison with other assessments.** Teachers should understand the importance of external validation, and be able to use several assessments of students' learning based on different assessment procedures, and compare performances on these to see if there are glaring discrepancies.

12. **Removing bias and prejudice in classroom assessments.** Item bias or prejudice that carries a negative reflection on particular subgroups may over the long run have a negative impact on these subgroups. Bias may also occur in the structure and design of assessments: what is chosen and emphasized. Teachers must recognize and control biases of both types in the materials they use, expectations they have of students, and how they report on and relate to students.

Teachers are not well equipped to design and develop assessments of process and higher-level cognitive skills, and often do not use systematic procedures to guide their performance assessments. This, as Stiggins (1988a) notes, is an instructional priority.

The reviewers generally suggested that theoretical development in this area is necessary, and that classroom prescriptions may be difficult. Instruction should emphasize meaningful content- and grade-differentiated examples of how assessments can be designed that are in keeping with important program goals and applicable in the classroom. This should be taught both in assessment courses and as part of training in pedagogy, and these courses must provide considerable practice to ensure that teachers are confident in applying the notions. Short courses are not enough, and teachers should have practice in real, school settings.

Recommendation 10 was also considered more important than three of the other recommendations. This recommendation concerns the structural aspect of validity: assessments must produce scores in keeping with the purpose (e.g., summative, diagnostic), and that these should reflect accurately the intended importance and weighting of constituent parts. It is usually accomplished through the use of tables of specifications, which teachers rarely use, but includes confirming that this weighting is maintained in any composite scoring procedures, particularly those formed for grading. Reviewers thought that both theoretical development and specific prescriptions were desirable, and that this could probably be done for education students in common, with some differentiation for subject areas. This could be dealt with in a separate course, as part of pedagogy, or as a short course.

Recommendation 12 deals with controlling for bias and prejudice in assessment, and was considered more important than one other recommendation. Reviewers thought it could be accomplished through some theory and specific classroom suggestions for all students in common. It was not thought to be difficult to achieve, and could be part of pedagogy, primarily through lectures. Clarification of what this entails and procedures to assist educators are readily available in sources such as Cole and Moss (1989), Shepard (1982), and Tittle (1982).

Recommendations 8, 9, and 11 received less support. Recommendation 8 deals specifically with diagnostics, which was not considered highly important at this point in teacher preparation, partly because of the level of training necessary to carry out diagnostics properly, and for teachers in the case studies, was not considered part of teaching social studies or science at the junior high level. It may be possible to provide instruction to prospective elementary and primary teachers in diagnostic assessment in selected areas, such as reading and mathematics. Recommendations 9 and 11 involve empirical approaches to validation and as with other technical aspects, were not considered essential to initial teacher preparation, as others have argued (e.g., Stiggins, 1991a).

## Utility and Efficiency

These two characteristics are considered together since efficiency can be thought of as an aspect of utility. Utility itself is properly thought of as part of the structural aspects and consequential nature of validity: the four recommendations for utility relate to how scores are obtained, interpreted, and communicated. The recommendation for efficiency, 17, deals with the practical problem of preparing teachers to optimize their assessment time, particularly since alternative assessment formats are being recommended by leading measurement and cognitive psychology researchers. These types of assessment may require assessment designs that differ considerably from the more traditional paper and pencil types, and rely heavily of performance appraisal, thus likely to require more teacher and classroom time. The five utility and efficiency recommendations are:

13. **Making value interpretations of assessment results using referencing procedures.** Teachers must understand the issue of score referencing relative to norms, criteria, or individuals, and be able to apply the relevant techniques in specific situations. Teachers must also understand that part of the task is setting standards, which involves professional judgement.

14. **Obtaining the scores necessary to use the assessment results properly.** Teachers must know how to design assessments so that appropriate scores and information can be obtained. The assessment must produce discrimination of scores at useful points on a scale if the scale is to be converted into grades or mastery-nonmastery categories.

15. **Communicating the results and interpretations of assessments.** Teachers must be able to communicate assessment results in variety of ways and to different audiences. This would include numerical results, summary statistics, grades, anecdotal reports, diagnoses, and affective evaluations.

16. **Controlling the effects of subjectivity in observing and marking.** Teachers must be skilled in using different procedures for subjective evaluations, such as rating schemes, holistic scoring, and narrative descriptions.

17. **(Efficiency) Making maximum use of assessment time and effort.** Assessment effort should be roughly proportional to the significance of the evaluation and its consequences. Teachers must know ways to speed up the procedures for applied performance assessment including observation, selection-type tests, and longer constructed-response assessments.

Recommendation 15 was rated as more important than three others, and recommendation 14 more important than one other. The remaining two utility recommendations, 13 and 16, and the one for efficiency, 17, were rated more important than one other recommendation as well, but were rated somewhat lower than 15. Recommendations 14 and 15 are related, 14 refers to obtaining appropriate scores and 15 concerns teachers' ability to communicate assessment results, which are topics that recently have received considerable interest among researchers (e.g., Brookhart, 1992; Frary, Cross, & Weber, 1992; Friedman & Manley, 1991; Manke & Loyd, 1991; Nava & Loyd, 1992). The reviewers thought that instruction could be common for all students, although there are clear elementary-secondary differences in grading and reporting practices. Anecdotal or descriptive reporting, for example, is now a mainstay of elementary schools (e.g., J. O. Anderson & Bachor, 1993; Brookhart, 1992; Friedman & Frisbie, 1993; Schulz, 1993; Webster, 1987) whereas letter and number grades are common at the secondary level (e.g., Loyd, Nava, & Hearn, 1991; Friedman & Manley, 1991; Stiggins, Frisbie, & Griswold, 1989).

Reviewers suggested both theoretical development of these topics and specific prescriptions for the classroom, which, at this point, would be based more on measurement experience and lore (e.g., Taylor, 1977) than on well-developed theory and research. There was slight favouring of presenting this in a course on pedagogy for recommendation 15, but not for 14. On the other hand, 14 is more of a technical matter, and may be best addressed in a course on measurement and evaluation. The lecture approach seemed most suitable.

Recommendation 13 deals with forming value interpretations based on assessments, which is a related to 14 and 15, and reviewers suggested that it may be best covered in pedagogy. This appears sensible since 13 deals with values and setting standards, which are grounded in curriculum as well as expectations and so forth. Recommendation 16 deals with observer subjectivity in scoring, which is also covered in recommendations 5 and 12, and may be best dealt with in the context of validity and subjectivity, as these relate to bias and fairness. Recommendation 17 is of lesser importance.

## Conclusions and Recommendations for Teacher Preparation

The recommendations for teacher preparation were reviewed and based on this a number of recommendations were rated as more important: recommendations 1, 2, 5, 6, 7, 10, 12, 14, and 15 (see above). This provides guidance for the focus and approach to instruction in classroom assessment for education students. It incorporates the practical realities of the classroom from analysis of the case studies, and applies the principles of measurement and evaluation through the identification of important characteristics of assessment. The possible content for a course in classroom assessment for prospective teachers can be inferred from the *Standards for Teacher Competence* (American Federation of Teachers et al. (1990; also Frisbie & Friedman, 1987) and *Principles for Fair Student Assessment Practices* (1993), but is documented in a number of sources (see Linn, 1990; Nitko, 1991a; Rogers, 1990b, 1991; Schafer, 1991). Topics based on the case studies are identified above.

The problem is determining from the vast array of what might be included in a course on classroom assessment, what is most defensible. Based on the characteristics of assessment, the recommendations noted above should further guide the choice of content. Only a small part of prospective teacher's undergraduate life will be devoted directly to the study of education, and a smaller part to assessment. This means that some aspects of assessment must be carefully selected as necessary to teach under a rubric separable from the broader concerns of curriculum and teaching. Other aspects should be incorporated into, or at least supported by, courses on pedagogy, either those directly concerned with specific curriculum in the schools or those dealing with teaching and learning. Some aspects are more relevant in some programs than others, and others may be necessary to differentiate for teachers of the lower and higher grade levels.

It is hard to argue against the importance of the topics outlined by Linn (1990) and Rogers (1991), but as Stiggins (1991a) cogently argues, some compromises must be made and these probably should be made in the interests of classroom realities. On this basis, some of the technical aspects of validity and reliability are necessary to forfeit, but it is probably not acceptable to ignore the problems associated with forming composite scores, setting standards and interpreting scores, and obtaining grades and other forms of evaluative statements, although these may be fully as technical and involved. The case studies and the review of recommendations provided just such a basis for compromise. The ratings of importance gave a measure of test of the recommendations. Potential topics were outlined earlier in this chapter, but the overall guiding principles are:

1. Teachers must receive instruction in determining what is to be assessed and what are the most appropriate ways in which assessment should be conducted. This requires both theoretical development and specific classroom suggestions. It should also be a joint venture between those with responsibility in measurement and those responsible for curriculum and teaching. Without this, it is difficult to give teachers sufficient base for actually developing and using assessments with varying formats that focus on such things as process skills and higher level thinking.

2. Teachers must receive instruction in designing assessments which can provide information for the purposes identified, including both formative (e.g., informing classroom instruction, individual diagnostics) and summative (e.g., grading and reporting). This includes designing assessments which are appropriately weighted in both content topics and skill levels, and that provide usable scores (e.g., meaningful composites, profiles).

3. Teachers must be skilled in designing and preparing paper and pencil assessments, both formal and informal. This remains one of the mainstays of assessment in classrooms, particularly at higher grade levels, and yet the quality of teacher-made tests has been shown to be poor. Teacher education must focus directly on the necessary skills, but also must permit ample practice within subject areas and at pertinent grade levels.

4. Teachers must be skilled in assessing students' products and behaviours. The importance of direct assessment, or performance assessment (and authentic assessment), is well-recognized, and teachers must be given ample background in how this can be done, and relevant practice based on actual program objectives and school situations.

5. Teachers must receive instruction in what affective objectives are generally considered important to teach and assess, and how these can be assessed meaningfully and fairly in a classroom assessment. Some of these are of a generic nature, and it is difficult to pinpoint who would have responsibility for identifying them (e.g., honesty, effort, cooperativeness, respect for others, tolerance of the views of others). It is important to note that many of these appear on student reporting forms used by schools, and this may serve as a starting point for identifying what affective characteristics prospective teachers should be able to assess. Others would be more linked to curricular areas such as science or social studies (e.g., valuing information, openness to new theories).

6. Teachers must be skilled in combining information from assessments and interpreting this information to students, parents, and others. There are many formats by which this can be done, and teachers must be able to use several of these. Grading and reporting these as letters or numbers is one set of such procedures, but there are others which are becoming more prevalent at the lower grade levels: for example, descriptive and anecdotal reporting. Teachers need assistance in performing this function. At present, there are no well-developed theoretical frameworks and supporting research to guide practice, but the topic must be addressed.

7. Teachers should understand the qualities of good assessment procedures, and the need for obtaining sufficient information to make important decisions. It is unreasonable to expect that teachers will begin to use technical procedures to evaluate their assessments, but they can become skilled in setting and applying criteria for various types of assessment. These have been clearly identified for paper and pencil tests (e.g., Gronlund & Linn, 1990), but less so for other formats. These should be developed for most if not all procedures used by teachers, but, more importantly, for procedures advocated by educators for application in schools, but that have not been as well developed: performance assessment, portfolio assessment, affective assessment, authentic assessment.

8. Teachers should have an understanding of the theoretical bases of measurement and evaluation, but this can be forfeited if it entails encroachment on development of the other seven principles. This basis should focus on issues of validity, including bias

and subjectivity, but more on how this impinges on classroom practice than on psychometrics. Issues of reliability are, at best, secondary as a technical exercise.

These eight guiding principles accommodate the higher-rated recommendations (listed above), but do not parallel them. For example, the first guiding principle accommodates validity recommendation 6: in deciding what to assess, teachers address the subject matter and identify the important learnings for their students, which are often higher-level cognitive skills. But this principle also includes aspects of validity noted in recommendation 7: the learnings to be assessed usually relates to a subject area, and this helps inform instruction in the area. The guiding principles also generally incorporate the competencies derived from the case studies, although this link cannot be made as simply. For example, competency 4 clearly is related to the first principle as well, but it is also related to principles 2, 3, 4, and 5.

Instruction related to the nine higher-rated recommendations corresponds to the principles in this way:

| | Guiding principle | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Higher-rated recommendations | | | | | | | | |
| Reliability: 1 | | | | | | * | * | * |
| 2 | | | * | * | * | | * | * |
| 5 | | | | * | * | * | * | |
| Validity: 6 | * | | * | * | | | * | |
| 7 | * | * | * | * | * | | | |
| 10 | | * | | | | * | * | * |
| 12 | | | * | * | * | * | | * |
| Utility: 14 | | * | * | * | * | * | * | * |
| 15 | | | | | * | * | * | |

The recommendations did not address when, or at what stage, in a teacher's education should assessment be learned. No best place or time for developing teachers' assessment skills emerged from the study. The research did not attempt to determine this as it focused on preservice teacher preparation.

## Implications for Further Research and Practice

There are a number of areas that require further study. It is important to extend this research to observe the assessment practices of larger numbers of teachers. This extension should involve teachers of different subject areas so that differences in assessment procedures and teacher needs could be documented. The present study involved teachers of science and social studies, and, while it is expected that there would be many similarities between them and teachers of other subjects, there are curricular differences that suggest that some of their assessment practices would vary. For example, it may be that teachers of the arts view their assessment tasks differently than do teachers of the sciences. Also, these two groups may require different types of assessment procedures.

This study involved teachers at the junior high level. There are suggestions in the literature that teachers' approaches to assessment differ at lower and at the higher grade levels. For example, there are clear differences in the reporting procedures used in elementary and high schools.

The study did not provide clear evidence as to how and where in the professional development of teachers preparation in classroom assessment should occur. However, some of the comments from the reviewers suggested that teachers should have background in assessment prior to becoming a teacher. There is also the view that teachers must have some understanding of children, and of instruction, before they can come to grips with the issues of assessment. Strong grounding in the subject matter is suggested by some; this argument has merit since teachers are expected to create their own assessments of complex and subtle skills germane to understanding in a subject domain. There is a clear need for research into developing assessments appropriate to a range of skills in a subject domain, and how these can be adapted to classroom use.

## REFERENCES

Ahmann, J. S., & Glock, M. D. (1981). *Evaluating pupil growth* (6th ed.). Toronto: Allyn & Bacon.

Aiken, L. R. (1976). *Psychological testing and assessment.* Boston: Allyn & Bacon.

Airasian, P. W. (1988a). Symbolic validation: The case of state-mandated, high-stakes testing. *Educational Evaluation and Policy Analysis, 10*(4), 301-313.

Airasian, P. W. (1988b). Measurement driven instruction: A closer look. *Educational Measurement: Issues and Practice, 7*(4), 702-709.

Airasian, P. W. (1991a). *Classroom assessment.*. Toronto: McGraw-Hill.

Airasian, P. W. (1991b). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice, 10*(1), 13-16, 26.

Airasian, P. W., Kellaghan, T., Madaus, G. F., & Pedulla, J. J. (1977). Proportion and direction of teacher rating changes of pupils' progress attributable to standardized test information. *Journal of Educational Psychology, 69*(6), 702-709.

Algina, J. (Ed.). (1992). The National Assessment of Educational Progress [Special issue]. *Journal of Educational Measurement, 29*(2).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1988). *Code of fair testing practices in education.* Washington, DC: American Psychological Association, Joint Committee on Testing Practices.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: American Federation of Teachers.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, J. O. (1987). *Teacher practices in and attitudes towards student assessment.* Paper presented to the annual meeting of the Canadian Educational Researchers' Association, Hamilton, ON.

Anderson, J. O. (1989). Evaluation of student achievement: Teacher practices and educational measurement. *The Alberta Journal of Educational Research, 35*(2), 123-133.

Anderson, J. O. (1990). Editorial: Assessing classroom achievement. *The Alberta Journal of Educational Research, 36*(1), 1-3.

Anderson, J. O, & Bachor, D. G.. (1993). *Assessment practices in the elementary classroom: Perspectives of the stakeholders.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Anderson, L. (1981). *Assessing affective characteristics in the schools.* Toronto: Allyn & Bacon.

Anderson, S. B., Ball, S., Murphy, R. T., & Associates. (1975). *Encyclopedia of educational evaluation.* San Francisco: Jossey-Bass.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Erlbaum.

Arter, J. (1993). *Designing scoring rubrics for performance assessments: The heart of the matter.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Barnes, S. (1985). A study of classroom pupil evaluation: The missing link in teacher education. *Journal of Teacher Education, 36*(4), 46-49.

Bateson, D. J. (1990). Measurement and evaluation practices of British Columbia science teachers. *Alberta Journal of Educational Research, 36*(1), 45-51.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1-17.

Berk, R. A. (1982a). Introduction. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 1-8). Baltimore, MD: John Hopkins.

Berk, R. A. (Ed.). (1982b). *Handbook of methods for detecting test bias.* Baltimore, MD: John Hopkins.

Berk, R. A. (1984a). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 231-266). Baltimore, MD: John Hopkins.

Berk, R. A. (Ed.). (1984b). *A guide to criterion-referenced test construction.* Baltimore, MD: John Hopkins.

Berk, R. A. (Ed.). (1986a). *Performance assessment: Methods and applications.* Baltimore, MD: John Hopkins.

Berk, R. A. (1986b). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* Toronto: Academic Press.

Block, J. H. (Ed.). (1971). *Mastery learning: Theory and practice.* New York: Holt, Rinehart & Winston.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. Handbook I: Cognitive domain.* New York: David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* Toronto: McGraw-Hill.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning.* Toronto: McGraw-Hill.

Boothroyd, R. A., McMorris, R. F., & Pruzek, R. M. (1992). *What do teachers know about measurement and how did they find out?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago: University of Chicago.

Boyd, J., Jacobsen, K., McKenna, B. H., Stake, R. E., & Yachinsky, J. (1975). *A study of testing practices in the Royal Oak (Michigan) public schools.* Royal Oak, MI: Royal Oak Public School District.

Brandt, R. (Ed.). (1985). The search for solutions to the testing problem [Special issue]. *Educational Leadership, 43*(2).

Brandt, R. (1989). On the misuse of testing: A conversation with George Madaus. *Educational Leadership, 46*(8), 26-29.

Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 292-334). Baltimore, MD: John Hopkins.

Brennan, R. L., & Kane, M. T. (1983). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*(2), 277-289.

Brookhart, S. M. (1992). *Teachers' grading practices: Meaning and value.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Brookhart, S. M. (1993). *Grading and classroom management: What does it mean to earn a grade?* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: Holt, Rinehart & Winston.

Burgess, R. G. (Ed.). (1985). *Strategies of educational research: Qualitative methods.* Philadelphia, PA: Palmer.

Burstall, C. (1986). Innovative forms of assessment: A United Kingdom approach. *Educational Measurement: Issues and Practice*, *5*(1), 17-22.

Campbell, S. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Canadian Education Association. (1989). *Grade promotion and retention: Practices in Canadian school boards.* Toronto: Author.

Carlberg, C. (1981). *South Dakota study report.* Denver, CO: Midcontinent Regional Educational Laboratory.

Carlson, S. B. (1985). *Creative classroom testing. Ten designs for assessment and instruction.* Princeton, NJ: Educational Testing Service.

Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education, 35*(6), 57-60.

Carter, K. (1986). Testwiseness for teachers and students. *Educational Measurement: Issues and Practice, 5*(4), 20-23.

Cartwright, C. A., & Cartwright, G. P. (1985). *Developing observation skills* (3rd ed.). Toronto: McGraw-Hill.

Chambers, B. A. (1982). *Quality control review of teacher-made tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Chambers, B. A., & Fleming, M. (1982). *Test screen: An instrument for rating teacher-made tests.* Cleveland, OH: Cleveland Public Schools.

Chase, C. I. (1978). *Measurement for educational evaluation.* Don Mills, ON: Addison-Wesley.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 255-296). New York: Macmillan.

Coffman, W. E. (1985). *Testing in the schools: A historical perspective.* Los Angelos: University of California, Center for the Study of Evaluation.

Cohen, S. A. (1987). Instructional alignment: Searching for the magic bullet. *Educational Researcher, 15*(8), 16-20.

Cohen, S. A., & Hyman, J. S. (1991). Can fantasies become facts? *Educational Measurement: Issues and Practice, 10*(1), 20-23.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36*, 1067-1077.

Cole, N. S. (1987). A realist's appraisal of the prospects for unifying instruction and assessment. In *Assessment in the service of learning* (pp. 103-117). Proceedings of the 1987 ETS Invitational conference. Princeton, NJ: Educational Testing Service.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 201-219). New York: Macmillan.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field studies.* Boston, MA: Houghton Mifflin.

Cousins, J. B. & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331-364.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Toronto: Holt, Rinehart & Winston.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*, 1-14.

Cronbach, L. J. (1977). *Educational psychology* (3rd ed.). New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

David, J. L. (1979). *Local uses of the Title I evaluations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Dawson-Sanders, B., Reshetar, R. A., & Shea, J. A. (1992). *Alterations to item text and effects on item difficulty and discrimination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Dorr-Bremme, D. W. (1983). Assessing students: Teachers' routine practices and reasoning. In UCLA Center for the Study of Evaluation, *Evaluation Comment, 6*(4), 1-12.

Dorr-Bremme, D.W., & Herman, J. L. (1984). *Testing and assessment in American public schools: Current practices and directions for improvement.* Los Angelos: University of California, Center for the Study of Evaluation.

Dorr-Bremme, D.W., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices.* Los Angelos: University of California, Center for the Study of Evaluation.

Doyle, R. (1989). The resistance of conventional wisdom to research evidence: The case of retention in grade. *Phi Delta Kappan, 71*(3), 215-220.

Ebel, R. L. (1967). Improving the competence of teachers in educational measurement. In J. Flynn & H. Garber (Eds.), *Assessing behavior: Readings in educational and psychological measurement,* (pp. 171-182). Reading, MA: Addison-Wesley.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2*(2), 7-10.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice items: What are textbook authors telling teachers? *Journal of Educational Research, 83*(5), 289-293.

Fairbairn, D. J. (1988). Pupil profiling: New approaches to recording and reporting achievement. In R. Murphy & H. Torrance (Eds.), *The changing face of educational assessment* (pp. 35-66). Philadelphia, PA: Open University.

Feiman-Nemser, S., & Floden, R. E. (1986). The cultures of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 505-526). New York: Macmillan.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105-146). New York: Macmillan.

Fennessy, D. (1982). *Primary teachers' assessment practices: Some implications for teacher training*. Paper presented at the annual conference of the South Pacific Association for Teacher Education, Frankston, Victoria, Australia.

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.). Toronto, ON: McGraw-Hill.

Fleming, M. (1979). *Classroom measurement needs in 1980's*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Fleming, M., & Chambers, B. A. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement,* no. 19 (pp. 29-38). San Francisco: Jossey-Bass.

Frary, R. B., Cross, L. H., & Weber, L. J. (1992). *Testing and grading practices and opinions in the nineties: 1890s or1990s?* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Frick, T., & Semmell, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research, 48*(1), 157-184.

Friedman, S. J., & Frisbie, D. A. (1993). *The validity of report cards as indicators of student performance.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Friedman, S. J., & Manley, M. (1991). *Grading practices in the secondary school: Perceptions of the stakeholders.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Frisbie, D. A. (1988). Reliability of test scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25-35.

Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4*(1), 67-83.

Frisbie, D. A., & Friedman, S. J. (1987). Test standards--Some implications for the curriculum. *Educational Measurement: Issues and Practice*, *6*(3), 17-23.

Frisbie, D. A., & Waltman, K. K. (1992). Developing a personal grading plan. *Educational Measurement: Issues and Practice, 11*(10), 35-42.

Fuchs, L. S., Deno, S., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*(2), 449-460.

Gable, R. K. (1986). *Instrument development in the affective domain.* Boston: Kluwer-Nijhof.

Gipps, C., & Goldstein, H. (1983). *Monitoring children: An evaluation of the Assessment of Performance Unit.* London: Heinemann.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237-262.

Goehring, H. J., Jr. (1973). Course competencies for undergraduate courses in educational tests and measurements. *The Teacher Educator, 9*(1), 11-20.

Goetz, J. P., & LeCompte, M. D. (1984). *Ethnography and qualitative design in educational research.* Orlando, FL: Academic Press.

Goslin, D. A. (1967). *Teachers and testing.* New York: Russell-Sage.

Government of Newfoundland and Labrador. (1990). *The evaluation of students in the classroom: A handbook and policy guide.* St. John's, NF: Department of Education, Government of Newfoundland and Labrador.

Green, K. E., & Stager, S. F. (1986). *Effects of training, grade level, and subject matter taught on the types of tests and test items used by teachers.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Griswold, P. A., & Griswold, M. M. (1992). *The grading contingency: Graders' beliefs and expectations and the assessment ingredients.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Gronlund, N. E. (1981). *Measurement and evaluation in teaching* (4th ed.). New York: Macmillan.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching* (5th ed.). New York: Macmillan.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Gullickson, A. R. (1982). *The practice of testing in elementary and secondary schools.* Paper presented at the 1982 Rural Education Conference at Kansas State University, Manhattan, KS.

Gullickson, A. R. (1984a). *Matching teacher training with teacher needs in testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Gullickson, A. R. (1984b). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77*(4), 224-246.

Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research, 79*(2), 244-248.

Gullickson, A. R. (1986a). *The characteristics of preservice teacher preparation in educational measurement: Professor perspectives.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Gullickson, A. R. (1986b). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23*(4), 347-354.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. *Educational Measurement: Issues and Practice, 4*(1), 15-18.

Gullickson, A. R., & Hopkins, K. D. (1987). The context of educational measurement instruction for preservice teachers: Professor perspectives. *Educational Measurement: Issues and Practice*, *6*(3), 12-16.

Gulliksen, H. (1986). Perspective on educational measurement. *Applied Psychological Measurement, 10*(2), 109-132.

Guskey, T. (1986). Staff development and the process of teacher change. *Educational Researcher, 15*, 5-11.

Haertel, E. (1986). *Choosing and using classroom tests: Teacher's perspectives on assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11*(1), 21-25.

Hall, B. W., Carroll, D., & Comer, C. B. (1988). Test use among classroom teachers and its relationship to teaching level and teaching practices. *Applied Measurement in Education, 1*(2), 145-156.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199-230). Baltimore, MD: John Hopkins.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*(3), 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70*(9), 683-687.

Hathaway, W. E. (1983). Editor's notes. In W. E. Hathaway (Ed.), *Testing in our schools: New directions for testing and measurement,* no. 19 (pp. 1-3). San Francisco: Jossey-Bass.

Haynes, S. N., & Wilson, C. C. (1979). *Behavioral assessment.* San Francisco, CA: Jossey-Bass.

Herman, J. L., & Dor-Bremme, D. W. (1983). Uses of testing in the schools: A national profile. In W. E. Hathaway (Ed.), *Testing in our schools: New directions for testing and measurement,* no. 19 (pp. 7-17). San Francisco: Jossey-Bass.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Merrill.

Hills, J. R. (1989). *Training priorities reflected in introductory measurement textbooks.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59*(3), 297-313.

Houts, P. L. (Ed.). (1977). *The myth of measurability*. New York: Hart.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-475.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jett, D. L., & Schafer, W. D. (1992). *Classroom teachers move to center stage in the assessment arena--ready or not!* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Johanson, G. A. (1992). *A compromise grading model for classroom tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Johnson, D. W., & Johnson, F. P. (1991). *Joining together: Group theory and skills* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Jolly, S. J., & Gramenz, G. W. (1984). Customizing a norm-referenced achievement test to achieve curricular validity: A case study. *Educational Measurement: Issues and Practice, 3*(3), 16-18.

Jones, R., & Carbol, B. (1988). *A summary of Canadian assessment practices*. Unpublished manuscript, British Columbia Ministry of Education, Vancouver, BC.

Kellaghan, T., Madaus, G.. F., & Airasian, P. W. (1980). *Standardized testing in elementary schools: Effects on schools, teachers, and students*. Washington, DC: National Institute of Education, Department of Health, Education and Welfare.

Kirkland, M. (1971). The effects of tests on students and schools. *Review of Educational Research, 41*(4), 303-350.

Kirst, M. (1991a). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20*(2), 21-23, 27.

Kirst, M. (1991b). Interview on assessment issues with James Popham. *Educational Researcher, 20*(2), 24-27.

Klopfer, L. E. (1971). Evaluation of learning in science. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning* (pp. 559-641). Toronto: McGraw-Hill.

Kunder, L. H., & Porwoll, P. J. (1977). *Reporting pupil progress: Policies, procedures, and systems*. (ERS Report), Arlington, VA: Educational Research Services.

Lanier, J. E., & Little, J. W. (1986). Research on teacher education. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 527-569). New York: Macmillan.

Lazar-Morrison, C., Polin, L., Moy, R., & Burry, J. (1980). *A review of the literature on test use*. Los Angelos: University of California, Center for the Study of Evaluation.

LeCompte, M. D., & Goetz, J. P. (1982). Problems of reliability and validity in ethnographic research. *Review of Educational Research, 52*(1), 31-60.

LeMahieu, P. G. (1984). The effects on achievement and instructional content of a program of student monitoring through testing. *Educational Evaluation and Policy Analysis, 6*(2), 175-187.

LeMahieu, P. G., & Leinhardt, G. (1985). *Higher test scores: Educational improvement or content manipulation?* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

LeMahieu, P. G., & Wallace, R. C. (1986). Up against the wall: Psychometrics meets praxis. *Educational Measurement: Issues and Practice, 5*(1), 12-16.

Ligon, G. D. (1983). Preparing students for standardized testing. In W. E. Hathaway (Ed.),*Testing in our schools: New directions for testing and measurement,* no. 19 (pp. 19-27). San Francisco: Jossey-Bass.

Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. *Teachers College Record, 91*(3), 422-436.

Lissitz, R. W., Schafer, W. D., & Wright, M. (1986). *Measurement training for school personnel: Recommendations and reality.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Livingstone, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph, 61*(4).

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3* (Monograph Suppl. 9), 635-694.

Lortie, D. (1975). *School teacher.* Chicago: University of Chicago Press.

Loyd, B. H., Nava, F. J. G., & Hearn, D. L. (1991). *High school students perceptions of the grading process.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

MacRury, K. (1988). *Assigning course grades: Policies and practices.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Madaus, G. F. (1981). Reactions to the "Pittsburgh Papers". *Phi Delta Kappan, 62*(9), 634-636.

Manke, M. P., & Loyd, B. H. (1990). *An investigation of non-achievement related variables in teachers' grading practices.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Manke, M. P., & Loyd, B. H. (1991). *A study of teachers' understanding of their grading practices.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Marso, R. N., & Pigge, F. L. (1988). *An analysis of teacher-made tests: Testing practices, cognitive demands, and item construction errors.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Marso, R. N., & Pigge, F. L. (1992). *A summary of published research: Classroom teachers' knowledge and skills related to the development and use of teacher-made tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education.* Toronto: Allyn & Bacon.

Mayo, S. T. (1964). What experts think teachers ought to know about educational measurement. *Journal of Educational Measurement, 1*(1), 79-86.

McKee, B. G., & Manning-Curtis, C. (1982). *Teacher constructed classroom tests: The stepchild of measurement research.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

McLean, L. D. (1985). *The craft of student evaluation in Canada.* Toronto: Canadian Education Association.

McLean, L. D. (1990). Time to replace the classroom test with authentic measurement. *The Alberta Journal of Educational Research, 36*(1), 78-84.

McMorris, R. F., & Boothroyd, R. A. (1992). *Tests that teachers build: An analysis of classroom tests in science and mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch. *Educational Measurement: Issues and Practice, 3*(3), 9-15.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practice, 8* (1), 14-22.

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology* (3rd ed.). Toronto: Holt, Rinehart & Winston.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Toronto: Holt, Rinehart & Winston.

Merwin, J. C. (1982). Standardized tests: One tool for decision making in the classroom. *Educational Measurement: Issues and Practice, 2*(3), 14-16.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20.

Messick, S. (1985). Progress and standards as standards for progress: A potential role for NAEP. *Educational Measurement: Issues and Practice, 4*(4), 16-19.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan.

Miller, H. G., Williams, R. G., & Haladyna, T. M. (1978). *Beyond facts: Objective ways to measure thinking.* Englewood Cliffs, NJ: Educational Technology.

Mulholland, L. A., & Berliner, D. C. (1992). *Teacher experience and the estimation of student achievement.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mullis, I. V. S. (1992). Developing the NAEP content-area frameworks and innovative assessment methods in mathematics, reading, and writing. *Journal of Educational Measurement, 29*(2), 111-131.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Nava, J. G., & Loyd, B. H.. (1992). *The effects of student characteristics on the grading process.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Neill, D. M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan, 70*(9), 688-697.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287-293.

Newman, D. C., & Stallings, W. M. (1982). *Teacher competency in classroom testing, measurement preparation, and classroom testing practice.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Nickerson, R. S. (Ed.). (1989a). Special issue on educational assessment [Special issue]. *Educational Researcher, 18*(9).

Nickerson, R. S. (1989b). New directions in educational assessment. *Educational Researcher, 18*(9), 3-7.

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research, 50*(3), 461-485.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction.* Don Mills, ON: Academic Press.

Nitko, A. J. (1991a). Editorial: What are we teaching teachers about assessment and Why?. *Educational Measurement: Issues and Practice, 10*(1), 2.

Nitko, A. J. (Ed.). (1991b). The practical matter of setting standards [Special issue]. *Educational Measurement: Issues and Practice, 10*(2).

Noll, V. H. (1955). Requirements in educational measurement for prospective teachers. *School and Society, 82*, 88-90.

Noll, V. H., Scannell, D. P., & Craig, R. C. (1979). *Introduction to educational measurement.* Boston: Houghton Mifflin.

Norris, S. P. (1989). Can we test validly for critical thinking. *Educational Researcher, 18*(9), 21-26.

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacifiic Grove CA: Midwest.

O'Sullivan, R. G. & Chalnick, M. K. (1991). Measurement related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*(1), 17-19, 23.

Oosterhof, A. C. (1987). Obtaining intended weights when combining students' scores. *Educational Measurement: Issues and Practice, 7*(4), 29-37.

Pilcher-Carlton, J., & Oosterhof, A. C. (1993). *A case study analysis of parents', teachers', and students' perceptions of the meaning of grades: Identification of descrepancies, their consequences, and obstacles to their resolution.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. *Phi Delta Kappan, 59*(9), 585-588.

Pipho, C. (1980). News from the states. *NCME Measurement News, 18*(5), 19-23.

Plake, B. S., & Berk, R. A. (1984). Development and analysis of survey instruments about NCME's annual meeting. *Educational Measurement: Issues and Practice, 3*(2), 40-41.

Plake, B. S., & Witt, J. C., (1986). The future of testing. *Applied Psychological Measurement, 10*(4), 405-414.

Popham, W. J. (1981). *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 29-48). Baltimore, MD: John Hopkins.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan, 68*(9), 679-682.

Popham, W. J. (1988). *Educational evaluation.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1990). *Modern educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12-15.

Popham, W. J. (1992). A tail of two test specification strategies. *Educational Measurement: Issues and Practice, 11*(2), 16-17, 22.

Power, C., & Wood, R. (1984). National assessment: A review of programs in Australia, United Kingdom and United States. *Comparative Education Review, 26*, 8-16.

Priestley, M. (1982). *Performance assessment in education and training: Alternative techniques.* Englewood Cliffs, NJ: Educational Technology.

*Principles for fair student assessment practices for education in Canada.* (1993). Edmonton, AB: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Proctor, D. (1985). Selecting a standardized test for a district-wide testing program. *Educational Measurement: Issues and Practice, 3*(3), 25-26.

Quellmalz, E. S. (1985). Needed: Better methods for testing higher-order thinking skills. *Educational Leadership, 43*(2), 29-36.

Quinto, F. (1977). Teacher-made tests--An alternative to standardized tests. *Today's Education, 66*(2), 52-53.

Resnick, D. P. (1980). Minimum competency testing historically considered. *Review of Educational Research, 8*, 3-29.

Resnick, D. P. (1981). Testing in America: A supportive environment. *Phi Delta Kappan, 62*(9), 625-628.

Resnick, L. B. (1981). Introduction: Research to inform a debate. *Phi Delta Kappan, 62*(9), 623-625.

Robinson, G., & Craver, J. (1989). *Assessing and grading student achievement. ERS report.* Arlington, VA: Educational Research Service.

Roeder, H. H. (1973). Teacher education curricula--Your final grade is F. *Journal of Educational Measurement, 10*(2), 141-143.

Rogers, W. T. (1990a). Current educational climate in relation to testing. *The Alberta Journal of Educational Research, 36*(1), 52-64.

Rogers, W. T. (1990b). *Educational measurement in Canada: Evolution or extinction?* President's invited address presented at the annual meeting of the Canadian Educational Research Association, Victoria, BC.

Rogers, W. T. (1991). Educational measurement in Canada: Evolution or extinction? *The Alberta Journal of Educational Research, 37*(2), 52-64.

Roid, G. H. (1984). Generating the test items. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 49-77). Baltimore, MD: John Hopkins.

Roid, G. H., & Haladyna, T. M. (1980). The emergence of an item writing technology. *Review of Educational Research, 50*(2), 293-314.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* Toronto: Academic Press.

Ross, J. A., & Maynes, F. J. (1985). Retention of problem-solving performance in school contexts. *Canadian Journal of Education, 10*(4), 383-401.

Rudman, H. E., Kelly, J. L., Wanous, D. S., Mehrens, W. A., Clark, C. M., & Porter, A. C. (1980). *Integrating assessment with instruction: A review (1922-1980).* (Research series No. 75). East Lansing, MI: Michigan State University, Institute for Research on Teaching.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *Phi Delta Kappan, 62* (9), 631-634.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation.* Belmont, CA: Wadsworth.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*(1), 3-6, 12.

Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57-63.

Schulz, H. W. (1985). *Summary of provincial assessment practices in Canadian public education.* Paper prepared for the Council of Ministers of Education, Canada, Toronto.

Schulz, H. W. (1993). *Grading and reporting practices in Newfoundland and Labrador schools.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Schwager, M. T., Mitchell, D. E., Mitchell, T. K., & Hecht, J. B. (1992). How school district policy influences grade level retention in elementary schools. *Educational Evaluation and Policy Analysis, 14*(4), 421-438.

Scriven, M. (1977). *Evaluation thesaurus.* Inverness, CA: Edgepress.

Shavelson, R. J. (1988). *Statistical reasoning for the behavioral sciences* (2nd ed.). Toronto, ON: Allyn and Bacon.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*(3), 447-465.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore, MD: John Hopkins.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, 47*(8), 4-9.

Shepard, L. A., & Bliem, C. L. (1993). *Parents' opinions about standardized tests, teacher's information and performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Shepard, L. A., & Smith, M. L. (Eds.). (1989). *Flunking grades: Research and policies on retention.* London: Falmer.

Shulman, L. S. (1980). Test design: A view from practice. In E. L. Baker & E. S. Quellmalz (Eds.), *Educational testing and evaluation: Design, analysis, and policy* (pp. 63-73). Los Angelos: Sage.

Sirotnik, K. A. (1983). *Toward more sensible achievement measuring: A view and review.* Los Angelos: University of California, Center for the Study of Evaluation.

Slavin, R. E. (1978). When does cooperative learning increase student achievement? *Psychological Bulletin, 94*, 53-63.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher, 20*(5), 8-11.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurment: Issues and Practice, 10*(4), 7-11.

Spradley, J. P. (1979). *The ethnographic interview.* New York: Holt, Rinehart & Winston.

Sproull, L., & Zubrow, D. (1981). Standardized testing from the administrative perspective. *Phi Delta Kappan, 62*(9), 628-631.

Stenhouse, L. (1988). Case study methods. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 49-53). Toronto: Pergamon.

Stetz, F. P., & Beck, M. D. (1979). *Comments from the classroom: Teachers' and students' opinions of achievement tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Stetz, F. P., & Beck, M. D. (1981). Attitudes toward standardized tests: Students, teachers, and measurement specialists. NCME: *Measurement in Education, 12*(1), 1-10.

Stiggins, R. J. (1985). Improving assessment where it means the most: In the classroom. *Educational Leadership, 43*(2), 69-74.

Stiggins, R. J. (1986a). *Lessons from the observation of classroom assessment environments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Stiggins, R. J. (1986b). *Evaluating students by classroom observation: Watching students grow.* Washington, DC: National Education Association.

Stiggins, R. J. (1987a). *Profiling classroom assessment environments.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Stiggins, R. J. (1987b). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(2), 33-42.

Stiggins, R. J. (1988a). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 69*(5), 363-368.

Stiggins, R. J. (1988b). *The nature and quality of teacher-developed classroom assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Stiggins, R. J. (1990). Toward a relevant classroom assessment research agenda. *The Alberta Journal of Educational Research, 36*(1), 92-97.

Stiggins, R. J. (1991a). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7-12.

Stiggins, R. J. (1991b). *A practical guide for developing sound grading practices.* Portland, OR: Northwest Regional Educational Laboratory, Center for Classroom Assessment.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, *22*(4), 271-286.

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice*, *5*(1), 5-17.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice*, *8*(2), 5-14.

Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26*(3), 233-246.

Stiggins, R. J., Rubel, E., & Quellmalz, E. (1988). *Measuring thinking skills in the classroom.* Washington, DC: National Educational Association.

Subkoviak, M. J. (1984). Estimating the reliability for mastery-nonmastery classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267-291). Baltimore, MD: John Hopkins.

Taylor, H. (1979). *Grading practices: Issues and alternatives.* Victoria, BC: Province of British Columbia, Ministry of Education.

Terwilliger, J. S. (1977). Assigning grades: Philosophical issues and practical recommendations. *Journal of Research and Development in Education, 10*(3), 21-39.

Terwilliger, J. S. (1987). *Classroom evaluation practices of secondary teachers in England and Minnesota.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Terwilliger, J. S. (1989). Classroom standard setting and grading practices. *Educational Measurement: Issues and Practice, 8*(2), 15-19.

Thayer, J. D. (1991). *Use of observed, true and scale variability in combining students' scores in grading.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston, MA: Houghton Mifflin.

Thorndike, R. L. (1988). Reliability. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 330-343). Toronto: Pergamon.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). Toronto: Wiley.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). *Measurement and evaluation in psychology and education* (5th ed.). Toronto: Collier Macmillan.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 31-63). Baltimore, MD: John Hopkins.

Tittle, C. K. (1988). Test bias. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 392-397). Toronto: Pergamon.

Traub, R. E., Nagy, P., MacRury, K., & Klaiman, R. (1988). Teacher assessment practices in high school calculus. In L. Pereira-Mendoza & M. Quigley (Eds.), *Proceedings of the 1989 Annual Meeting, Canadian Mathematics Education Study Group*. St. Catherines, ON: Brock University.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10* (1), 37-45.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *Review of Research in Education* (vol. 9) (pp. 377-435). Washington, DC: American Educational Research Assocication.

Tuckman, B. W. (1975). *Measuring educational outcomes: Fundamentals of testing*. New York: Harcourt Brace Jovanovich.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26*(2), 191-208.

Wainer, H., & Braun, H. (Eds.). (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Waltman, K. K., & Frisbie, D. A. (1993). *Parents' understanding of their children's report card grades*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Wanous, D. S., & Mehrens, W. A. (1981). Helping teachers use information: The data box approach. NCME: *Measurement in Education, 12*(4), 1-10.

Ward, J. G. (1980). Teachers and testing: A survey of knowledge and attitudes. In L. M. Rudner (Ed.), *Testing in our schools*. Washington, DC: National Institute of Education.

Webster, J. B. (1987). *Teacher initiated assessment: A report for River East School Division*. Unpublished manuscript, River East School Division, Winnipeg, MB.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 81-129). Washington, DC: American Council on Education.

White, K. R., & Carcelli, L. (1982). *The effect of item format on computation subtest scores of standardized tests*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Wiersma, W., & Jurs, S. G. (1985). *Educational measurement and testing*. Toronto: Allyn & Bacon.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Toronto: Allyn & Bacon.

Wiggins, G. (1989a). Teaching to the (authentic) test. *Educational Leadership, 46*(8), 41-47.

Wiggins, G. (1989b). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.

Wilson, R. J. (1989). Evaluating student achievement in an Ontario high school. *Alberta Journal of Educational Research, 35*(2), 134-144.

Wilson, R. J. (1990). Classroom processes in evaluating student achievement. *Alberta Journal of Educational Research, 36*(1), 4-17.

Wilson, S. M., & Hiscox, M. D. (1984). Using standardized tests for assessing local learning objectives. *Educational Measurement: Issues and Practice, 3*(3), 19-22.

Wise, A., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1984). *Teacher evaluation: A study of effective practices.* Santa Monica, CA: Rand Corporation.

Wittrock, M. C., & Baker, E. L. (Eds.). (1991). *Testing and cognition.* Englewood Cliffs, NJ: Prentice-Hall.

Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational Leadership, 46*(8 ), 35-39.

Wolf, D. P., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Research in Education* (vol. 17) (pp. 31-74). Washington, DC: American Educational Research Association.

Womer, F. (1970). National assessment says. NCME: *Measurement in Education, 2*(1), 1-7.

Wood, P., Bennett, T., & Wood, J. (1990). *Grading and evaluation practices and policies of school teachers.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Yager, R. E. (1989). Assess all five domains of science. *The Science Teacher, 54*(7 ), 33-37.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*(8), 10-14.

Yeh, J. P. (1978). *Test use in schools.* Los Angelos: University of California, Center for the Study of Evaluation.

Yeh, J. P., Herman, J. L., & Rudner, L. M. (1981). *Teachers and testing: A survey of test use.* Los Angelos: University of California, Center for the Study of Evaluation.

# APPENDICES

# APPENDIX A

## LETTERS RELATED TO THE CASE STUDIES

1. Letter to assistant superintendent requesting nominations of teachers.
2. Letter to assistant superintendent thanking the division for its support.
3. Letter indicating how the research information was to be used and safe guarded.

March 9, 1988

Dr. --------
Assistant Superintendent
--------
--------

Dear Dr. ---------:

I have reviewed my research project and have identified what I require in terms of interviewing and observing teachers from your division.

**Teachers to be interviewed and observed.**

I require four teachers, two teachers from each of two junior high schools:

- -two teachers of Social Studies and two teachers of Science--grade 7, 8, or 9;
- -preferably, this would be one science and one social studies teacher from each school (and it would be ideal if the two teachers in one school taught some of the same students).

I would like to interview each teacher twice, once prior to the classroom observations, and once during the span of the classroom observations or following the observations (times would be arranged at the convenience of the teacher and school):

- -the first interview would be introductory and informal with the intent to become familiar with the teacher, school, and particular program and course, and to determine the schedule for observations--I anticipate that this would require less than one-half (1/2) hours;
- -the second interview would be to collect information on such things as the purposes of various assessment activities and how they are used in the classroom, how grading is accomplished and assessments combined, etc., and to collect samples of assessment materials from the teacher--this likely would take one-half to three-quarters (1/2-3/4) hours.

I would like to observe each teacher several (3-4) times for approximately one class period each time over a period of 3 to 4 weeks (this would be at a time acceptable to the teacher and the school), but one time should be shortly before a teacher-made test was to be administered, or shortly after. I would not interfere with the teacher and class activities.

I can begin any time after April 2, 1988, and would like to proceed as soon as is reasonable.

The above are what I would like. However, there is considerable flexibility. Contact me directly at either of the numbers below if there are problems.

The purpose of the study is to determine the evaluation practices of classroom teachers, and to develop a model and procedures for the effective preparation and professional development of teachers in evaluation. For your information, I have enclosed an outline of the study, which includes a brief description of the two objectives that pertain to the collection of information from teachers (I can forward the full proposal if you wish):

**Objective 2.1.** Observe and interview a small number of exemplary teachers and describe in detail their assessment practices, and

**Objective 2.3.** Survey by questionnaire a larger group of similar teachers to support these descriptions.

As I mentioned on the phone I would very much like a copy of Janet Webster's report. It appears to relate directly to my work.

My address for the next two weeks is:

H. Schulz
Division of Educational Research Services
Faculty of Education
3-104 Education Building North
University of Alberta
Edmonton, AB
T6G 2G5

Phone: (403) 432-3762 (Office), 438-6790 (Home)

I shall be in Winnipeg and available to begin work on Tuesday, April 5, 1988. I will have a Winnipeg phone number as of that date: 257-1458, and my address will be: 68 Woodlawn Ave., Winnipeg R2M 2P2.

Thank you very much for your support and assistance.

Sincerely,

Henry Schulz

encl.: Classroom Assessment Practices

May 19, 1988

Dr. --------
Assistant Superintendent
--------
--------

Dear Dr. ---------:

I am presently completing my research data gathering with the two teachers from each of -------- and -------- Schools. As well as providing me with the information I required, it is a tremendously rewarding experience for me! It is a real pleasure to be in the classroom again, and particularly in the presence of competent and effective teachers. The cooperation I am receiving is well beyond what I could have anticipated. The principals supported what I am doing and recommended several teachers. These teachers agreed with my request, readily invited me into their classrooms, and afforded me with considerable interview time. Thanks very much for allowing me the opportunity to carry out my research in these schools. I am sending a note of thanks to the particular teachers whom I observed and to the principals of the two schools.

You requested that I provide you with a letter reiterating the purpose of the study and specifying how the data is to be safeguarded. This is contained in a separate letter (attached). If there is anything that I may have omitted, or you wish clarified, please let me know. I have included both my Winnipeg address and where I can be contacted in Edmonton (I will be at the University of Alberta for the month of June).

I am enclosing a copy of the attached letter with the letters to the teachers and principals so that they are fully aware of what is happening.

Thank you again for your assistance, and thanks also for sending me a copy of Dr. Webster's report.

Sincerely,

Henry Schulz
68 Woodlawn Ave.
Winnipeg, Manitoba
R2M 2P2

Phone: 257-1458

Division of Educational Research Services
3-104 Education Building North
University of Alberta
Edmonton, Alberta
T6G 2G5
Phone: (403) 432-3762

May 19, 1988

Dr. --------
Assistant Superintendent
--------
--------

### Research on Classroom Assessment Practices

During the months of April and May of 1988 I conducted research on the student assessment practices of classroom teachers in two schools in the Division. I received permission for this research from you and from the principals and teachers involved. The research consisted of my observing the teachers on several occasions and interviewing them in some detail on how they assess in their classes. The results of the research will be used by me as part of my doctoral dissertation.

**Purpose of the study**

The overall purpose of the study is to develop a model and procedures for the preparation and professional development of teachers in evaluation. The first part of the study is to describe in detail the assessment practices of several exemplary classroom teachers. These descriptions, and the more general information we have on the assessment practices of teachers, will form the basis for identifying what evaluation procedures are practicable in the classroom and what ought to be included in the model.

**Observing and interviewing teachers**

Over a period of six weeks in April and May of 1988, two teachers in -------- School and two in -------- School were observed in approximately six classroom periods each. Each of the teachers was also interviewed after the observations; this involved two or three actual interview sessions. Times for the observations and interviews were arranged to suit the schedule of the teachers and the schools.

**Anonymity and reporting**

The observations and interviews provide a basis of data from which to describe teacher assessment practices. This will be a summary description. In the description no teacher will be identified, and every effort will be made so that any reports based on the study cannot attribute particulars of the findings to specific schools and teachers.

A report will be made to the Division; it will contain a summary description. As well, the schools involved are offered a presentation by the researcher on the study and its findings.

It has been made clear to the participants that the descriptions and the findings of the study will not be available until the next school year.

**Reference to the Division**

It is expected that no direct reference to the Division need be made in any reporting on the study. However, in the event that this may be necessary, clearance will be sought from the Superintendent's office of the Division.

Henry Schulz, Doctoral candidate
Department of Educational Psychology
University of Alberta
Edmonton, Alberta

(on educational leave from Manitoba Education)

# APPENDIX B

## FORMS USED TO OBTAIN TEACHERS' STRUCTURED RESPONSES TO EACH OF THE SIX ASPECTS RELATED TO CLASSROOM ASSESSMENT

I. Assessment Purposes

II. Assessment Methodology. . . achievement

III. Assessment Methodology. . . affective traits

IV. Assessment Selection Criteria

V. The Teacher and Assessment Sources

VI. The Teacher and Assessment Time

## I. Assessment Purposes

***K. Relative importance of purposes: Indicate the relative importance to YOU of these purposes for student assessment .***

Below are defined **ten** different purposes or decision areas that classroom assessments could serve (A to J), and a scheme for determining the relative importance awarded them by classroom teachers. Given **100 points** to distribute across these **10 purposes**, how would you distribute them to show the relative importance of the various decisions listed? The more points you assign to a purpose the more important it is to you.

| 10 Purposes | Approximate points out of 100 |
|---|---|
| A. Diagnosing individual needs of students. Identifying student strengths and weaknesses | |
| B. Diagnosing group needs. Detecting common instructional needs across a group of students. | |
| C. Assigning grades. Determining letter grades for report cards as feedback to students and parents | |
| D. Grouping for instruction. Subdividing a class into smaller instructional groups on assessment results. | |
| E. Identifying students for special services. Selection for placement into advanced or remedial program. | |
| F. Controlling and motivating students. Using assessment or the prospect of assessment to cause students to behave in a specific way, such as using assessment for motivation or for punishment | |
| G. Evaluating instruction. Documenting the success or failure of a particular instructional treatment. | |
| H. Communicating achievement expectations. Informing students of the nature of the content and skills they are to learn. | |
| I. Communicating affective or behavioural expectations. Informing students of the attitudes, values, preference, and behavioural patterns that are acceptable | |
| J. Providing test taking experience. Familiarizing students with item types and test conditions to prepare them for future tests. | |

Comments:

## II. Assessment Methodology

***A-12 Relative importance of methods: Indicate the relative importance to YOU of these methods for student assessment of achievement (for all purposes).***

Below are defined eleven different assessment methods for the assessment of student achievement (1 to 11), and a scheme for determining the relative importance awarded them by classroom teachers. Given **100 points** to distribute across these **11 methods**, how would you distribute them to show the relative importance of the various methods listed? The more points you assign to a method the more important it is to you.

Approximate
11 Assessment Methods points out of 100

1. Teacher -developed paper and pencil tests and quizzes: Multiple-choice, true-false, matching, fill-in, and long answer/essay tests **designed and written by the teacher**. ........................................ ______
2. Text-embedded paper and pencil tests and quizes: Multiple-choice, true-false, matching, fill-in, and long answer/essay tests **provided by the text publisher**. ........................................ ______
3. Performance assessment: Measurement via observation of student products and behaviours and evaluation via professional judgement....... ______
4. Oral questioning strategies: Questions asked by the teacher during instruction. ........................................ ______
5. Standardized tests: School, division, province-wide or program related assessments based on large-scale administration of published tests........ ______
6. Group assessment methods: Assessments in which students work together for a group grade or an individual grade.......................... ______
7. Opinions of other teachers: Positive or negative feelings about a student's achievement expressed by colleagues verbally or in written records. ........................................ ______
8. Assessment of reasoning skills: Measuring student's thinking skills through the application of Bloom's taxonomy or some equivalent structure ........................................ ______
9. Regular homework assignments: Periodic assignments designed to provide practice and yield information on student performance............. ______
10. Student peer ratings: Students rate each others performance................ ______
11. Student self ratings: Students evaluate their own performance............. ______

Comments:

________________________________________________

## III. Assessment Methodology

***B-10 Relative importance of methods for assessing affective traits: Indicate the relative importance to YOU of these methods for student assessment (for all purposes).***

Below are defined eight different assessment methods for the assessment of student affective traits (1 to 8), and a scheme for determining the relative importance awarded them by classroom teachers. Given **100 points** to distribute across these **8 methods**, how would you distribute them to show the relative importance of the various methods listed? The more points you assign to a method the more important it is to you.

**8 Affective Assessment Methods** — Approximate points out of 100

1. Observing individual students: Inferring affective traits from the observation of the behaviour of individual students.......................... _______
2. Observing group interactions: Inferring affective traits from observation of the social and academic interactions among students and between student and teacher. ............................................................ _______
3. Using questionnaires: Paper and pencil instruments used to gather affective data. ................................................................... _______
4. Using interviews: Formal and informal one on one oral exchanges of information between teacher and students(s) to gather affective data....... _______
5. Opinions of other teachers: Comments about student affect obtained from colleagues verbally or via past student records. ........................ _______
6. Opinions of other students: Comments about student affect obtained from other students............................................................ _______
7. Opinions of parents: Comments about student affect obtained from the student's parent or guardian.................................................. _______
8. Past student records: Draw inferences regarding affective characteristics from information obtained in the student's cumulative record. ............. _______

Comments:

_______________________________________________

_______________________________________________

## IV. Assessment Selection Criteria

### J. *Relative importance of criteria: Indicate the relative importance to YOU of these criteria for selection of assessment method.*

Below are defined nine different criteria for the selection of methods for the assessment of student achievement (A to I), and a scheme for determining the relative importance awarded them by classroom teachers. Given **100 points** to distribute across these **9 criteria**, how would you distribute them to show the relative importance of the various criteria listed? The more points you assign to a criterion the more important it is to you.

**9 Criteria for the Selection of Assessment Methods** — Approximate points out of 100

A. Results of method fit purpose of assessment: The chosen assessment method promises to fit the teacher's information needs...................... _______

B. Method matches intended outcomes: Match between the assessment format and the student characteristic measured.............................. _______

C. Ease of development: Amount of time, effort and technical skill required to use assessment method.......................................... _______

D. Ease of scoring: Amount of time, effort and technical skill required to score assessments based on this method................................. _______

E. Origin of assessment: Person originally responsible fro developing the assessment instrument and/or procedures. ................................... _______

F. Time required to administer the assessment: Testing units administered administered per unit of time. ................................................ _______

G. Degree of objectivity: Amount of teacher judgement that goes into the scoring process. .............................................................. _______

H. Applicability to measuring higher order thinking skills: Extent to which assessment method can serve to measure more that the recall of knowledge.................................................................... _______

I. Effective control of cheating: Contribution of method to test security and/or student copying....................................................... _______

Comments:

_______________________________________________

_______________________________________________

## V. The Teacher and Assessment Sources

### *A-2. Sources of assessment knowledge: Relative contribution of various sources to YOUR knowledge of assessment methodology.*

Below are defined seven different sources of knowledge about assessment of students (a to g), and a scheme for determining the relative importance awarded them by classroom teachers. Given **100 points** to distribute across these **7 sources**, how would you distribute them to show the relative importance of the various sources listed and those you may add? The more points you assign to a source the more important it is to you.

| 7 Sources of Assessment Knowledge | Approximate points out of 100 |
|---|---|
| a. Preservice and graduate teacher training. | ______ |
| b. Inservice training programs. | ______ |
| c. Ideas and suggestions from colleagues. | ______ |
| d. Readings from professional literature. | ______ |
| e. Guidebooks accompanying texts. | ______ |
| f. Own classroom experience. | ______ |
| g. Other (specify): |  |
| ________________________________________ |  |
| ________________________________________ | ______ |

Comments:

________________________________________

________________________________________

## VI. The Teacher and Assessment Time

***B-2. Relative amount of time spent in assessment procedures: The manner in which YOU use your time in the procedures of assessment.***

Below are defined seven different aspects of the assessment process (a to g), and a scheme for determining the relative amount of time awarded them by classroom teachers. Given **100 percentage points** to distribute across these **7 aspects**, how would you distribute them to show the relative amount of time you devote to each? The more points you assign to an aspect the more time you give to it.

Approximate percentage points out of 100

7 Aspects of the Assessment Process

a. Reviewing and selecting assessments.................................................______

b. Developing own assessments.................................................______

c. Administering assessments.................................................______

d. Scoring assessments.................................................______

e. Recording results.................................................______

f. Providing feedback.................................................______

g. Evaluating assessment quality.................................................______

Comments:

______________________________________________________________

______________________________________________________________

# APPENDIX C

## RATINGS OF TEST ITEM QUALITY IN TEACHER-MADE TESTS

### Choice Format

True-False Items

Multiple-Choice Items

Matching

Other Alternate-Choice Items

### Short-Answer Format

Short Answer

Completion/Fill-in-the-Blank/Identification/Association

### Essay Format

Restricted Response
Extended Response

### Context-Dependent Format

Objective Interpretive Exercise
Short Answer
Essay

Rating of Test Item Quality (by Item Type)

| Criteria | Rating scale[a] 1 | 2 | 3 | 4 | Number of tests containing item type |
|---|---|---|---|---|---|
| **Choice Format** | | | | | |
| **True-False Items** (also T-F correction, yes-no, fact-opinion, correct inference-incorrect inference, etc.) | | | | | |
| 1. Item type most appropriate | 2 | 4 | 3 | - | 9 |
| 2. Items clearly true or false | 6 | 3 | - | - | 9 |
| 3. Clear, simple language | 9 | - | - | - | 9 |
| 4. T & F statements similar length | 8 | 1 | - | - | 9 |
| 5. Approximately equal number of T's and F's | 8 | 1 | - | - | 9 |
| 6. No trick/trivial statements | 3 | 5 | 1 | - | 9 |
| 7. No specific determiners | 7 | 2 | - | - | 9 |
| 8. No negatives/double negatives | 8 | 1 | - | - | 9 |
| 9. No detectable T-F pattern | 9 | - | - | - | 9 |
| **Multiple Choice** | | | | | |
| 1. Item type most appropriate | 5 | 4 | - | - | 9 |
| 2. Stem has clear, central problem | 8 | 1 | - | - | 9 |
| 3. Stem free of irrelevant material | 9 | - | - | - | 9 |
| 4. Stem pos., or neg. highlighted | 9 | - | - | - | 9 |
| 5. Choices grammatically consistent | 8 | 1 | - | - | 9 |
| 6. Choices brief, no irrelevancies | 7 | 2 | - | - | 9 |
| 7. Choices similar in length, form | 8 | 1 | - | - | 9 |
| 8. Choices in "natural" order | 7 | 1 | - | - | 8 |
| 9. Choices free of verbal clues | 8 | 1 | - | - | 9 |
| 10. Very few "none/all of the above" | 9 | - | - | - | 9 |
| 11. Answer is clearly correct or best | 6 | 3 | - | - | 9 |
| 12. Answer location has no pattern | 8 | 1 | - | - | 9 |
| 13. Plausible distractors | 2 | 5 | 2 | - | 9 |

[a] 1=Nearly all items (>85%), 2=Most items (50-85%), 3=Some items (15-49%), 4=Few items (<15%), ?=Cannot be determined

| **Criteria** | **Rating scale**[a] **1** | **2** | **3** | **4** | Number of tests containing item type |
|---|---|---|---|---|---|
| **Matching** | | | | | |
| 1. Exercise type most appropriate | - | 5 | 3 | - | 8 |
| 2. Homogeous material in each list | 1 | 3 | 3 | 1 | 8 |
| 3. Unequal list lengths | 6 | - | 1 | 1 | 8 |
| 4. Lists of reasonable length (5-10) | 4 | 1 | - | 3 | 8 |
| 5. Lists in "natural" order | - | 2 | - | - | 2 |
| 6. Responses brief | 5 | 1 | 2 | - | 8 |
| 7. Responses on right-hand side | 5 | 1 | - | 2 | 8 |
| 8. Responses used more than once | 1 | 1 | - | 4 | 6 |
| 9. Directions give basis of matching | 2 | 2 | 3 | 1 | 8 |
| 10. Where/how to answer is given | 4 | 4 | - | - | 8 |
| 11. Complete exercise on one page | 7 | - | - | 1 | 8 |
| **Other Alternate-Response** (e.g., Key response) | | | | | |
| 1. Exercise type most appropriate | 3 | 1 | - | - | 4 |
| 2. Responses appropriate to all stems | 3 | 1 | - | - | 4 |
| 3. Reasonable number of responses | 3 | 1 | - | - | 4 |
| 4. Responses in "natural" order | 4 | - | - | - | 4 |
| 5. Responses brief | 4 | - | - | - | 4 |
| 6. Directions give basis of response | 4 | - | - | - | 4 |
| 7. Where/how to answer is given | 4 | - | - | - | 4 |
| 8. Complete exercise on one page | 4 | - | - | - | 4 |
| **Short-Answer Formats** | | | | | |
| **Short Answer** | | | | | |
| 1. Item type most appropriate | 9 | 11 | 2 | - | 22 |
| 2. Items are clear, direct questions | 16 | 4 | 2 | - | 22 |
| 3. Items free of language, etc. clues | 20 | 2 | - | - | 22 |
| 4. Answer is brief phrase, etc. | 17 | 5 | - | - | 22 |
| 5. Item has only one correct answer | 10 | 10 | 2 | - | 22 |
| 6. Degree of precision/units given-- | 1 | 4 | - | - | 5 |
| 7. Relevant content only is marked | ? | ? | ? | ? | - |

[a] 1=Nearly all items (>85%), 2=Most items (50-85%), 3=Some items (15-49%), 4=Few items (<15%), ?=Cannot be determined

| **Criteria** | **Rating scale**[a] **1** | **2** | **3** | **4** | Number of tests containing item type |
|---|---|---|---|---|---|
| **Completion/Fill-in-the-blank Identification/Association** | | | | | |
| 1. Item type most appropriate | 5 | 17 | 1 | - | 23 |
| 2. Nontrivial information tested | 6 | 14 | 2 | 1 | 23 |
| 3. Not verbatim textbook language | 17 | 6 | - | - | 23 |
| 4. Items free of language, etc. clues | 20 | 3 | - | - | 23 |
| 5. Desired answer is clear | 14 | 6 | 3 | - | 23 |
| 6. Response blanks of equal length | 20 | 2 | 1 | - | 23 |
| 7. Blanks at end of statement | 16 | 3 | 3 | - | 22 |
| 8. Degree of precision/units given | 2 | 2 | 1 | - | 5 |
| **Essay Formats** | | | | | |
| **Restricted Response Extended Response** | | | | | |
| 1. Item type most appropriate | 12 | 3 | - | - | 15 |
| 2. Measure higher-level outcomes | 7 | 3 | 1 | 1 | 12 |
| 3. Specifies overall purpose | 14 | 1 | - | - | 15 |
| 4. Specifies content to be included | 6 | 7 | 2 | - | 15 |
| 5. Specifies form, structure, length | 3 | 8 | 4 | - | 15 |
| 6. Specifies sufficient time limits | 1 | - | - | - | ? |
| 7. Specifies marking guide | - | - | 9 | 3 | 12 |
| 8. Only relevant outcomes marked | ? | ? | ? | ? | ? |
| 9. Same items for all students | 11 | - | - | 4 | 15 |
| **Context-Dependent Formats** | | | | | |
| **Objective Interpretive Exercise Short Answer Essay** | | | | | |
| 1. Item type most appropriate | 4 | 4 | 2 | - | 10 |
| 2. Material relevant to outcomes | 7 | 2 | - | - | 9 |
| 3. Material appropriate to students | 7 | - | - | - | 7 |
| 4. Non-verbal material used | 8 | - | - | 1 | 9 |
| 5. Material somewhat novel | 3 | 2 | 1 | 2 | 8 |
| 6. Material brief, clearly interpretable | 4 | 4 | 1 | - | 9 |
| 7. Items require material | 9 | 1 | - | - | 10 |
| 8. Several items for material | 8 | 1 | - | 1 | 10 |
| 9. Items efficient in type/format | 9 | - | - | - | 9 |
| 10. Items meet relevant criteria | 2 | 8 | - | - | 10 |

[a] 1=Nearly all items (>85%), 2=Most items (50-85%), 3=Some items (15-49%), 4=Few items (<15%), ?=Cannot be determined

# APPENDIX D

## DOCUMENT PRESENTED TO REVIEWERS OF THE MODEL

Cover letter

Teacher Preparation in Classroom Assessment: Review of Recommendations

- Model for teacher preparation in classroom assessment
- Summary of recommedations for teacher preparation
- Teacher preparation and ongoing professional development
- RELIABILITY: Review of recommendations 1-5
- VALIDITY: Review of recommendations 6-12
- UTILITY: Review of recommendations 13-16
- EFFICIENCY: Review of recommendations 17

Henry W. Schulz
23 Walwyn Street
St. John's, Newfoundland A1A 3W5
(709) 579-7347

Faculty of Education
Memorial University of Newfoundland
St. John's, Newfoundland A1B 3X8
(709) 737-3502 FAX: (709) 737-2345

June 7, 1991

Dr. ----------
Department of ----------
----------
----------

I have prepared 17 recommendations that pertain to various aspects of teacher preparation and professional development in classroom assessment. Classroom assessment refers to all the various ways in which teachers assess their students in classrooms, for whatever reasons. The purpose of the study is to develop a model for teacher education in this area, particularly at the preservice level. The recommendations are based on a survey of the classroom assessment literature and on case studies involving several teachers of junior high science and social studies. These recommendations also reflect present theory and thinking in measurement and evaluation.

The model should reflect what are considered good classroom assessment practices, but it must also present a realistic and practicable approach to the development of prospective and practicing teachers. The recommendations are submitted for review by individuals who have experience with classroom assessment from a variety of perspectives. This includes practicing teachers, school and district/division administrators, curriculum specialists, and educational measurement specialists. You are identified as one of approximately 20 educators, who are being asked to provide input.

Could I ask that you review each recommendation in detail and give your comments. The focus for the review is given on pages 12-18 of the document, and the procedures that I suggest you use are outlined on page 19. One page is provided for your response to each recommendation (pages 20-36). The recommendations are summarized on pages 1-12 of the document.

I realize that this may take some time and effort on your part, and I very much appreciate this. The only recompense I can offer is a summary of the work when it is completed (but I hope the subject is of interest to you). Thanks in advance for your assistance.

Your responses can be forwarded to **A. Russell, P.O. Box 43, Moosehorn, Manitoba, R0C 2E0,** if you can do it within the next two weeks, else it can be left with Dr. Tom Maguire of the Centre for Research in Applied Measurement and Evaluation, Faculty of Education, 3-104.

Sincerely,

Henry W. Schulz
Doctoral candidate
Department of Educational Psychology
University of Alberta

This study is being conducted as part of my doctoral program, under the supervision of Dr. T. Maguire.

encl.: Teacher Preparation in Classroom Assessment: Review of Recommendations for the Model

# TEACHER PREPARATION IN CLASSROOM ASSESSMENT: REVIEW OF RECOMMENDATIONS FOR THE MODEL

**Henry W. Schulz**

Doctoral Student

Educational Psychology, University of Alberta

---

## CONTENTS

**Model for Teacher Preparation in Classroom Assessment**......................xxx

**Summary of Recommendations for Teacher Preparation**.........................xxx
  Reliability.........................................................................................xxx
  Validity............................................................................................xxx
  Utility...............................................................................................xxx
  Efficiency..........................................................................................xxx

**Teacher Preparation and Ongoing Professional Development**..................xxx
  Features for Review of Initial Education in Classroom Assessment......................xxx
  Procedures to Review the Recommendations for Teacher Preparation....................xxx

**RELIABILITY: Review of Recommendations 1-5**................................xxx

**VALIDITY: Review of Recommendations 6-12**...................................xxx

**UTILITY: Review of Recommendations 13-16**...................................xxx

**EFFICIENCY: Review of Recommendation 17**....................................xxx

## Model for Teacher Preparation in Classroom Assessment

A model outlining the professional preparation and development of teachers for classroom assessment must be both practical in the classroom and consonant with good assessment principles. Four major tasks are required to identify the model:

1. Specify the characteristics of good classroom assessment practices.
2. Evaluate present teacher assessment practices with respect to these characteristics.
3. Describe the features of the model.
4. Evaluate the model for its practicality to the classroom setting and its adherence to modern assessment principles.

Present teacher assessment practices are based on a review of the literature and on the case studies conducted for the study. Four teachers, who were identified as experienced and competent by their principals and superintendents, were selected for the case studies. Two of these teachers taught science and the other two taught social studies. The case studies consisted of observing the teachers in their classrooms periodically over two months. The focus of the observations was to determine the nature of their assessment practices. The teachers were then interviewed regarding their assessment practices. The assessment materials used by these teachers for one reporting period were gathered and analyzed.

Four general characteristics of classroom assessment were identified on the basis of modern assessment theory and present classroom practice. For each characteristic, recommendations for the appropriate preparation of teachers were presented and discussed. The recommendations are summarized in the next section. From this, the model for teacher preparation in classroom assessment is to be formulated. The basis for reviewing the recommendations for the model is presented on pages 15-18. The procedures for the review are outlined on page 19, and the materials to which response is requested are on pages 20-36.

## Summary of Recommendations for Teacher Preparation in Classroom Assessment

The characteristics of good classroom assessment can be described under four major headings, Reliability, Validity, Utility, and Efficiency. Within these broad categories 17 recommendations for teacher education in classroom assessment were identified. The recommendations are summarized below. They form the focus of teacher preparation for classroom assessment.

### Reliability

Reliability refers to the consistency of assessment results, to the stability of scores obtained from using assessment procedures. This implies that the overall procedures for an assessment should be included in considering the reliability. For example, if some form of composite scores are to be used in determining the grouping of students in a class, all of the procedures that are included in calculating the scores should be reviewed for their reliability. There are five recommendations regarding reliability. Some of these refer to the overall process by which classroom assessments are determined, whereas others refer to various aspects of this process.

### 1. The importance of reliability for high-stakes assessments.

The importance of reliability is directly dependent on the consequences of the assessment. In classroom assessment, reliability should be greatest for those assessment results having the most impact on the students (i.e., high-stakes assessments). Therefore, teachers must be able to identify the main purpose of an assessment and its implications for the student, both the personal and social consequences.

This is necessary also for secondary purposes if these exist. Simply put, if the purposes of the assessment have considerable consequences to students (or even to what happens in the classroom or school), then care must be taken to ensure reliability. To understand the importance of reliability for high stakes assessments, teachers must be aware of the implications that low reliability might have for various decisions.

The resource costs of making all classroom assessments procedures highly reliable are considerable, and well beyond what is reasonable to expect in schools. Thus, it is necessary for teachers to make some tradeoff between the reliability of an assessment and its efficiency. The most significant consequences for classroom assessments are usually related to grading and reporting of student progress, with promotion-retention decisions having the most dramatic effects on students. Reliability is particularly important for assessments used to inform these types of decisions (these are high-stakes assessments). Assessments for these purposes require greater effort on the part of teachers to ensure their reliability.

### 2. Practical ways of improving reliability in the classroom.

The reliability of assessment information can be enhanced in two general ways which are practical in the classroom: by making the assessment procedures more explicit and systematic, and by increasing the amount of high-quality information gathered. Teachers must know how these apply to particular classroom assessment practices. Teachers must also know how to enhance the reliability of important types of assessment: subjective observations, constructed response testing, objective testing, portfolio assessment. Each of these has its unique problems, but the reliability of each can be enhanced.

The first principle implies that assessment procedures should be clearly understood by students so that their behaviours are not haphazard, and thus inconsistent. The assessment procedure must also be applied systematically to all students and to all occasions. For example, a student rating procedure must be understood so that extraneous clues do not lead to varied behaviour and inconsistent ratings. Furthermore, the rating procedures must be systematic so that similar behaviours are considered under similar conditions.

It is not sufficient to simply increase the "amount" of assessment included in providing information for important decisions. For example, just including the results of more classroom assignments will not necessarily increase the reliability of the combined data. Many of the assignments could suffer from the same flaws that diminish the reliability of any one. In the same way, the length of a test is no clear indicator of its reliability. The addition of further poor-quality items, to which student responses may be haphazard, will not improve the overall reliability of the test. For this principal to operate, the increased information must at least have some inherent reliability.

### 3. Increasing reliability through multiple quality assessments.

Students must be given the opportunity to exhibit their skills on several occasions, particularly if there is some doubt if the skill is mastered. This does not mean that "more is better" in terms of assessment, but that judicious choices of assessment procedures should be made for consequential evaluations. The reliability of scores is not enhanced by simply increasing the number of scores or assessments, but rather by increasing the occasions on which students can exhibit their learning. Teachers must be able to obtain assessments of skills and knowledge in several different ways.

This recommendation is an extension of the second part of recommendation 1, but emphasizes the importance of obtaining information on students on several occasions, and under differing conditions, if the skills are considered important enough to be assessed for some high-stakes purpose. This is particularly necessary in cases where it is not clear as to whether the skills are learned/mastered or not, or where scores are near a cutoff point of some kind. Judicious choices of assessment procedures should be made for the most consequential evaluations. As noted in the previous recommendation, it does not mean that "more is better" in terms of assessment. The reliability of scores is not enhanced by simply increasing the number of scores or assessments, but rather by increasing the appropriate occasions on which students can exhibit their learning. Teachers must be able to obtain assessments of skills and knowledge in several different ways to ensure that the underlying capability is being assessed accurately.

The general principle is that there ought to be multiple assessment procedures employed whenever an important decision is to be made. These procedures ought to be sufficiently varied so that extraneous skills do not override the capability of interest. For example, if only written-assignments are used to determine a student's understanding of scientific concepts, students with poor skills in writing may provide data that is incorrect, or inconsistent at best.

### 4. Numerical procedures for estimating reliability in classroom assessments.

There are a number of numerical procedures that can be used to help determine if classroom measurements are reliable. Some of these are fairly simple to compute, and, with modern computing equipment being readily available, it is becoming reasonable for teachers to check the reliability of their tests periodically. Teachers ought to confirm the adequacy of their measurements and check the reliability of some of the more critical assessments using straightforward numerical procedures.

There are several fairly simple statistics which are applicable a variety of scoring systems. The simplest of these is $p_0$ which is the proportion of consistent categorizations of students using two parallel assessment procedures (e.g., two items, two tests, two assignments). There are also programs available on microcomputers to compute a variety of test statistics, such as means, standard deviations, and correlations. It may be necessary to use split-half reliability procedures since frequently there is only one assessment. Teachers must know how to obtain and use some of these methods.

### 5. Reliability of subjective forms of assessment.

Assessments that involve subjective judgments on the part of teachers are an important, integral part of teaching. "Authentic" assessment often involves observing students applying their skills in actual situations (Wiggins, 1989a, 1989b), and therefore good observation techniques must be used. Even simulated situations, such as those designed for the classroom or the laboratory, often require direct observation and subjective judgement on the part of the teacher. Teachers must understand that subjective forms of assessment are prone to unreliability and to personal bias. They must also know practical procedures which can ameliorate these problems.

Prospective teachers must be able to develop various procedures that enhance the reliability of subjective evaluations. These procedures include such things as identifying observation and rating schemes, designing systematic schedules, recording observations, and conducting consistency checks on the procedures. Teachers must have the opportunity to practice subjective evaluations under a variety of conditions.

## Validity

Validity is the most important characteristic of assessment. It refers to the "correctness" of the results of an assessment. The unified view of validity, which is generally accepted today, implies that both logical and empirical evidence must be amassed, and that the consequences of the assessment must also be considered. The validation process has substantive, structural, and external components, and all of these are applicable to classroom assessment. No longer is content or curricular validation considered sufficient.

There are seven recommendations relating to validation of classroom assessment and teacher preparation. These encompass the three components of validity, but highlight aspects which are of particular importance to in the classroom yet practical.

### 6. Increasing assessment of important, higher level cognitive learnings.

Classroom assessment procedures must provide information on the learning that is most important for our students, which includes more complex and higher-level thinking. Therefore, teachers must have the knowledge and skills to design and conduct classroom assessments that are most appropriate to this kind of outcome. This includes both the content and thinking processes identified in subject areas, and the broader thinking skills, communications skills, and cognitive strategies (metacognition) that are presently emphasized by our schools.

The important learnings for students in schools have been clearly identified through curriculum guides, teaching materials, and professional documents and journals. These have not been unanimously agreed to, but there is general agreement that they include complex forms of learning. These learnings are often difficult to assess, so teachers must have adequate preparation in the development of appropriate assessments. Preparation would include general skills in developing assessments to tap higher-level thinking abilities and affective learnings, as well as skills to produce assessments particular to various subject disciplines. As an example, general skills would include such things as preparing instructions for students and developing scoring procedures for written work, whereas the language arts discipline would include learning to use analytic and descriptive scoring of prose passages.

### 7. Assessment that supports and informs good instruction.

Classroom assessment must reflect good pedagogy, both to support the learning process in the classroom as a mode of mental development by emphasizing what is important and setting standards for its attainment, and to provide practice for this kind of thinking (e.g., thinking skills that integrate learnings from several content topics or areas, and that impinge on future learning). Teachers must have training and actual experience in preparing and using assessment procedures that support the most important learning in classrooms, those that should be emphasized in class and on the assessments. Teachers must know how to develop and use a wide variety of assessment procedures. These include paper-and-pencil techniques, but more importantly, they should include systematic observation and applied performance assessment.

This recommendation extends upon the previous one. Not only must assessment emphasize that which is most important to learn, but it must also reflect what happens in the learning. Both the focus of the assessment and the style of the assessment must approach that which formed the learning. Teachers must use assessment procedures and formats that are in keeping with the instruction and the goals of the learning. This means that they must have a repertoire of assessment skills, including those necessary for the direct observation of behaviour, for analyzing student products, and for setting assignments and tests. Further, teachers must be able to focus the assessment on the intended learnings, and not to include extraneous factors.

### 8. Assessment for various purposes including instructional diagnostics.

Classroom assessment must be designed in such a way that it can provide the correct information for the intended purpose. Teachers must have the training to devise assessment procedures for a variety of purposes, such as for grading of students, individual and group instructional diagnosis, and evaluation of instruction. There are other purposes, but these are likely to be secondary. Teachers must also know when and how it is possible to accommodate more than one purpose, and to recognize when this leads to incompatible procedures.

Teachers need to know how to produce and use both formative and summative assessment procedures. These often require different approaches. Assessment for diagnosis requires that there be detailed information on meaningful diagnostic categories (skills, abilities), whereas summative assessments may sample from a number of content and skill domains. Although it is not practical to prepare teachers for individual diagnosis, or even instructional diagnosis, in all subject areas, teachers should understand the general principals and procedures, and be able to apply them in selected settings.

### 9. Empirical checks of validity of the skills covered in classroom assessments.

It is important for classroom assessment that several tasks or items devised to assess the same skill or concept yield similar results with students. A check for this presumed homogeneity is one way of testing if there is a meaningful, stable underlying construct, and that subtest scores can be interpreted. Without this, the construct can have little utility in future learning. Teachers must know how to check for homogeneity. This involves identifying the assessment tasks or items which should "hang together" and checking student results to see if this is the case.

This recommendation extends upon the previous one by specifying that teachers know procedures by which they can empirically check whether the assessment does provide diagnostic information, and that scores can be interpreted with some confidence for the purposes intended. For example, it is not enough to simply argue that several items in a test which appear to tap the same construct can be combined to obtain a measure of that construct. These items should correlate with one another, and this can be checked by looking at the performance of each item in relation to the other items assigned to the same subtest.

### 10. Making assessment results reflect the focus and purpose of the assessment.

Classroom assessments must be designed to provide the scores and information for which they are intended and these scores must be appropriate reflections of the importance of the content. Teachers must be trained to obtain scores from assessments corresponding to the nature and purpose of the assessment. This involves identifying clearly in advance of producing the assessment the approximate importance, and hence the weighting, of various areas of content and levels of skills (essentially producing a table of specifications for the assessment).

Recommendation 10 refers specifically to the procedures by which scores are obtained, either from direct ratings, from judgements based on observations or products, or from composites of assessment items. The scores must reflect clearly the main intent of the assessment, and should not be subverted by such procedures as inappropriate weighting of items or test components. This is particularly important in summative evaluation where composite scores are obtained from a variety of assessment sources. Procedures that are useful in this regard are the careful specification and weighting of aspects to be included in the assessments (e.g., test blueprints), and the explicit formulation of how composites are to be obtained.

### 11. Confirming the results of assessments by comparison with other assessments.

Classroom assessments must be designed to permit confirmation or disconfirmation of findings from previous assessments or from external sources. Previous assessments or other relevant evaluative information provide a check on assessment results, thereby giving external evidence for their validity. Teachers must base their judgements of students on information from multiple, diverse sources. The validity of the information should also be checked periodically in this way. This can be done informally by comparing the results of particular students on one assessment to those on another. It is preferable to use something more systematic, such as correlation, but it is unlikely that this will be done in the classroom setting. Teachers should understand the importance of external validation. They should attempt to obtain several assessments of students' learning using different assessment procedures, and compare students' performances on these to see if there are glaring discrepancies.

Teachers are clearly aware of the need to obtain multiple measures of student learning for making high-stakes decisions. But they should also determine if these multiple measures provide supportive data, which means there ought to be some consistency (convergence) in the scores. It is possible with low-cost computing equipment to conduct periodic checks for major assessments. The procedures could include preparing cross-tabulations and computing correlations.

### 12. Removing bias and prejudice in classroom assessments.

Classroom assessments should be relatively free of biases and prejudices that may have negative impact on students. There are two kinds of bias in assessment. One that may have direct deleterious effects on the assessment results of particular groups of students by virtue of their differences in experience which makes them perform below their actual skill level. This is the issue of fairness or equity. The other is more subtle, and may not have a direct effect on results, but it is bias or prejudice that has a negative reflection on particular subgroups and may over the long run have a negative impact on these subgroups. Teachers must be trained and experienced in recognizing biases of both types in the materials they use, in the expectations they have of students, and in how they relate to students. There are both judgmental and empirical procedures that can be invoked to reduce bias.

There are systematic procedures that can be used to remove some of the bias in assessment that are appropriate to the classroom assessment. Teachers should be aware of the common forms of bias that are reflected in stereotypes of subgroups based on characteristics such as gender, ethnicity and race, socioeconomic status, and geographical location. Guidelines exist in the preparation of educational materials that attempt to reduce potential biases, and teachers should be experienced in use of these guidelines. Teachers can also submit their assessments to peer and public review. However, more subtle forms of bias can exist in such things as language structure and complexity, and forms of response required of students. To reduce potential bias of this nature the performance of identified subgroups can be compared across various techniques of assessment. This highlights the need to use several methods in the accumulation of information for high-stakes decisions.

## Utility

The ability of assessments to provide information that is readily and appropriately interpreted is described under the topic of utility. This includes such aspects as the referential basis for the assessment scores, the clarity of the communication, and the objectivity of the scoring process. There are four recommendations under this heading.

### 13. Making value interpretations of assessment results using referencing procedures.

The interpretation of scores, or other assessment information, is a difficult and value-laden problem. There are no clear, simple guidelines that can be given. It is useful to think of score referencing in terms of norm-, criterion-, and self-referenced approaches to the problem. There are occasions where one approach is superior to another. Prospective teachers must understand the issue of score interpretation in terms of referencing, and be able to apply the relevant techniques in specific situations. In some cases criterion-referencing is the most appropriate, whereas in others it may be necessary to relate scores to the group or some other source of norms. Teachers must also understand that part of the task is setting standards. In some assessments the standards are imbedded in the assessment itself, but there is almost always opportunity for professional judgement. This professional judgement can be enhanced by including the views of others, and by making the process more public.

The interpretation of scores, and the reporting of value statements based on these interpretations, are imbedded in the context of the evaluation process. The standards of performance or the norms of behaviour are determined by such factors as the subject matter and the purpose of the educational program, the policies and context of the school and school system, and the personal and professional views of the individual teacher. Teachers must be able to explicate their standards and how they make value interpretations. This can then be understood by students (at least generally), and reviewed by others, such as parents and other teachers.

### 14. Obtaining the scores necessary to use the assessment results properly.

Test scores and their interpretations must be of use for the purposes intended. This means that the appropriate kinds of assessment results and interpretations must be obtainable from the assessment. A simple example of this is subtest scoring (to obtain a profile) versus total test scoring. Teachers must learn how to design the assessment so that the appropriate information is forthcoming. The assessment must produce discrimination of scores at useful points on a scale if the scale is to be converted into grades or mastery-nonmastery categories.

The problem of dealing with scores that are close to important cutoff points was discussed under recommendation 3. However, teachers should be aware of the potential problem in advance and devise assessments which have the potential of discriminating at important points on the scale (e.g., at scores near 80 when 80 and above is awarded an "A"). One way to do this with paper-and-pencil testing is to design items with the appropriate range of difficulty levels.

### 15. Communicating the results and interpretations of assessments.

Communication of assessment results is an important part of evaluation. Communication can have a number of audiences, each with different expectations and different abilities to understand the results. It is necessary to simplify interpretations of assessment results in many cases. For example, parents are confounded by too much information if they are confronted by their child's scores on all the objectives in a course. However, this may be very useful information to the teacher and student. It may also be necessary to explain clearly to an audience exactly how assessments have been interpreted and what this means to them. Teachers must be able to communicate assessment results in variety of ways and to different audiences. This would include numerical results, summary statistics, grades, anecdotal reports, diagnoses, and affective evaluations.

The most common form of formal reporting in schools are student report cards. These vary greatly in form and content from school to school and system to system. There is no one best way of reporting student learning, but there are suggestions of procedures are more effective than others. Prospective teachers must know what is generally considered important to report and ways of reporting that are intelligible to students and parents. Teachers must also know how to communicate directly with parents on their children's progress, such as in parent-teacher interviews.

### 16. Controlling the effects of subjectivity in observing and marking.

Both objective and subjective forms of assessment are necessary in classroom assessment, so teachers must be skilled in both. Teachers can be trained to make some of the subjective assessment procedures more systematic, if not entirely objective. Observations of student behaviors can be designed so that all students are observed under somewhat similar circumstances, for example. The goal is to enhance the quality of subjective assessments, and not to avoid them. Teachers must be trained to use a number of different procedures for subjective evaluations, such as rating schemes, holistic scoring, and narrative descriptions. These procedures can be grounded in subject areas, where particular approaches may be prominent.

The problems of reliability in subjective scoring are discussed under recommendation 5. The problem that is highlighted here is that or effectively transferring subjective judgements into statements that can be understood clearly. Often these statements are in the form of numbers, such as in holistic scoring of writing. But there are other ways that qualitative judgements can be communicated with clarity. Teachers must be familiar with ways of conducting subjective assessments and reporting them effectively. Education specialists working in subject area disciplines can provide guidance in this regard (e.g., how written work can be described and judged, how students' appreciation of art is determined).

## Efficiency

The last recommendation deals specifically with the relative cost in time and effort of conducting assessments of various kinds and for different purposes.

### 17. Making maximum use of assessment time and effort.

Since both teacher preparation time and classroom instruction time are at a premium, it is incumbent on teachers to be as efficient as possible in conducting their assessments. The amount of assessment effort on the part of the teacher and of the students should be roughly proportional to the significance of the evaluation and the consequences of the decision. Teachers must know the approximate effort and time required for various assessment procedures. They must also know ways to speed up the process. There are three general categories of assessment procedures that should be understood in this regard, applied performance assessment including observation, selection-type tests such as multiple choice, and longer constructed response assessments such as written papers and research reports.

There are a number of techniques that can be used by teachers to reduce the amount of time it takes to produce, administer, and score various types of assessment. For example, it is well recognized that good multiple-choice tests are difficult to produce, but quick to score. Word processing machinery greatly facilitates this process since assessment materials can be stored and readily modified. Many schools have access to computing equipment. Teachers should be trained in the use of commonly available equipment and applications.

## Teacher Preparation and Ongoing Professional Development in Classroom Assessment

Evidence from observations of teachers in the classroom (e.g., Stiggins, 1986; the case studies) and from surveys of teachers (Gullickson, 1986; Stiggins & Bridgeford, 1985) clearly implies that teachers do not typically apply the principles of good assessment procedures in their classrooms. Other evidence suggests that teachers do not well understand many of these principles (e.g., Chambers, 1982; Newman & Stallings, 1982; Ward, 1980), although some of the teachers in these studies had previous training in "tests and measurements". The importance of classroom assessment is readily apparent from the amount of effort teachers devote to it--estimates are in the neighbourhood of 20% and more of total classroom time. From this, there appears to be need to revise and extend initial teacher education preparation in classroom assessment, and also to provide ongoing continuing education of teachers with relevant inservice work.

There are a number of ways in which teachers can become better equipped for their classroom assessment tasks. Teacher development can be categorized as initial education, preservice or precertification, and continuing education, inservice professional development (Lanier & Little, 1986). Some of the knowledge and skills teachers require to adequately conduct assessment can be taught as part of initial teacher education, but probably much of this must be developed during the time a teacher is actually practicing in the classroom.

Those responsible for teacher education programs, teacher educators, must determine whether the focus is to prepare teachers in broad, theoretical bases of major disciplines of knowledge and their applications in education, or to provide them with the procedural skills and techniques of the teaching craft. Lanier and Little (1986) describe this as distinguishing the approach to teacher education as liberal-professional or technical-professional. Both approaches acknowledge the importance of perceiving teacher education as professional, which includes preparation related to the ethos and culture of the profession (Lortie, 1975), rather than purely academic, which emphasizes the subject discipline approach as commonly exemplified in faculties of arts and sciences. The liberal-professional approach emphasizes the "intellectually deep and rigorous study" of education from various vantage points having their methodological and substantive roots in disciplines such as the arts and sciences, including history, philosophy, and sociology. The technical-professional approach favours training of teachers in prescriptive knowledge and skill performance, these being based on process-product research evidence (outside-expert oriented) and on "tried-and-true" classroom techniques (authority oriented, and conformist). Lanier and Little (1986) argue that at present teacher education programs tend to the technical-professional approach, and conclude that:

> The increasing proportion of career teachers makes the oft-repeated call for a liberal-professional approach to teacher education all the more persuasive. . . . preparing career teachers for their continuing education requires greater emphasis on liberal-professional studies than is presently the case. . . . Unfortunately, changes in the teacher education curriculum have tended to move it in the opposite direction, giving increased dominance to the mastery of skills with immediate practical value. What is worse, studies of the curriculum of initial and continuing teacher education show it to be fragmented and shallow. (p. 555)

There is no one best approach to teacher education. However, there are issues of liberal versus technical education of teachers, of general knowledge versus specific skills, and of integrating the liberal aspects of teacher education with those of a professional character. This has implications for any program that purports to develop teachers' classroom assessment skills: for example, should teachers be taught the underlying theory and principles of assessment, with their attendant complexities, or should teachers be taught a narrow band of directly applicable assessment techniques?

It may not be practical, let alone reasonable, to expect prospective teachers to spend a large amount of their university time in developing an in-depth understanding of the broad field of measurement and evaluation, much of which would not be applicable to their future teaching situations. The amount of preservice time devoted to direct study in education is usually only a fraction of the overall initial education of teachers. Initial education generally consists of four or five years of formal coursework in general-liberal studies, in major and minor teaching areas, and in pedagogical study. It is estimated in the United States that "the course work in pedagogical studies generally represents only about one-fifth of a secondary teacher's required program and about one-third of an elementary teacher's program" (Lanier & Little, 1986, p. 529). Requirements differ considerably from one university to another and among programs within a given university, such as between a four-year BEd program and certification after a first degree. As an example, at the University of Alberta for 1990/91, of the 120 course weight requirements for a BEd degree 48 (40%) must be chosen from outside the Faculty of Education for the elementary route and approximately 54 (45%) from outside for the secondary route. However, some of the course requirements within the Faculty of Education would be considered foundational and not directly pedagogical (termed "basic education"); courses such as those in Educational Foundations and Educational Psychology, which may account for as many as 30 (25%) course weights of an elementary BEd program. Furthermore, a student may be certifiable if she/he obtains a degree from another faculty and completes 30 (25%) course weight requirements in Education. Initial teacher education in classroom assessment must be considered relative to the total amount of time in the teacher preparation program and to the amount of time that can be devoted to educational studies.

Although there are many aspects of classroom assessment which are appropriate to all levels and speciality areas for which teachers are prepared, it is unlikely that there would be much motivation for all students of education to be extensively trained in a common set of specific assessment skills. For example, one technique of considerable direct utility to teachers of language arts is holistic and analytic scoring of a variety of forms of student writing. However, students interested in teaching the sciences at higher grade levels could not be expected to use these skills to any great extent, and what may be of greater relevance to them would be learning how to effectively assess laboratory skills. What is more likely to be useful for all teachers is to be familiar with, but not extensively trained in, a number of procedures for the scoring of written material, including the application of selected procedures and their particular strengths and weaknesses. It would be desirable, however, for students in particular teacher training areas, such as language arts, to receive more extensive experience in this method of evaluation. It could be expected, then, that some aspects of assessment should be part of initial education for all prospective teachers whereas other components should be made specific to the anticipated teaching areas of particular groups of students.

Present thinking suggests that teacher socialization, as professionals, is much more a fact of teacher experience in schools than that of initial education (Feiman-Nemser & Floden, 1986). These authors further conclude that there is no one generalizable school culture, and variations exist in different schools and among different subgroups within a school. This implies that it would be difficult, and perhaps futile, to impart much of the understanding of the culture of teaching in initial teacher education. It also implies that some of the modes of teaching and norms of practice are developed while teachers are in the classroom. This suggests that at least some of the detailed practices pertinent to teaching in particular contexts would be more a matter of continuing education than of preservice education. Initial teacher education may only reasonably hope to achieve the more "liberal" goals of education.

A number of solutions have been proposed for enhancing teacher skills, both preservice (e.g., Gullickson, 1985) and inservice (e.g., Stiggins, 1987). However, the emphasis here is to identify a general model for the initial education of teachers, although reference can be made to possible continuing education where this appears appropriate.

## Features for Review of Initial Education in Classroom Assessment

There are a number of important characteristics of classroom assessment which teachers, irrespective of their specialities, ought to understand, and much of this content can be taught prior to experience in the classroom. This would include certain skills, and there are some issues and concerns teachers should be familiar with prior to teaching. Some of these would be suitable to teach to groups of teachers who are heterogeneous in their backgrounds and interests. Other areas may not be as applicable overall. These may require a differentiated approach in teacher education.

It is presumed that the prospective teachers have adequate substantive knowledge and understanding in their chosen specialty, if they are subject specialists. It is further assumed that they have the knowledge and appropriate background for the grade levels of interest, as well as an understanding of principles related to curriculum, learning, child development, and classroom organization and management.

The characteristics of good classroom assessment described earlier, reliability, validity, utility, and efficiency, serve as categories for the content and objectives of teacher education in assessment. The seventeen recommendations pertaining to these characteristics serve to guide the focus of this content, although the content is treated generally.

Four features of teacher education are identified which relate to the focus, organization, and delivery of initial education in classroom assessment. These features relate to the above comments on teacher education generally. These serve as categories for reviewing the recommendations for teacher preparation in classroom assessment. They are not discrete, and suggestions regarding one feature may have implications for the others.

1. Amount of substantive and theoretical development, and level of specific technical prescriptiveness of the instruction. This distinguishes between the substantive background to evaluation theory and methodology (e.g., mathematical and statistical underpinnings of measurement) and the actual suggested practices for classroom teachers (e.g., holistic scoring).

2. Type and level of differentiation of the material and its delivery, considering such characteristics as specialization of the prospective teachers: grade and age level (e.g., early, middle, senior years; adult learners); subject and discipline specialty (e.g., natural sciences, mathematics and related

fields, social sciences, humanities, languages and language arts, fine arts); education specialty programs (e.g., special education, gifted learners, mentally handicapped, counselling and support services).

3. Method and timing of delivery of the instruction, and relationship to the overall teacher education program: for example, separate course, short or minicourses related to specific topics within the broader program and courses of teacher education, integration of material into other teacher education courses, timing of delivery within the teacher education program, downward and upward articulation with other aspects of the teacher and specialist education program (and also programs of faculties outside education).

4. Nature of the learning experiences, such as lecture/discussion, laboratory work, clinical practice, and practical experience.

Not all four features need apply to each recommendation. Comments can be made for the recommendations with respect to the features. This is depicted as a matrix below.

**Model for Reviewing Teacher Preparation in Classroom Assessment**

| | | Features for reviewing the characteristics | | | |
|---|---|---|---|---|---|
| Characteristics of Classroom Assessment | Topics and Content of Instruction[a] | 1. Level of Specificity | 2. Common or Differentiated | 3. Method of Delivery | 4. Nature of Instruction |
| Reliability | Rec. 1-5 | ___ | ___ | ___ | ___ |
| Validity | Rec. 6-12 | ___ | ___ | ___ | ___ |
| Utility | Rec. 13-16 | ___ | ___ | ___ | ___ |
| Efficiency | Rec. 17 | ___ | ___ | ___ | ___ |

[a] The 17 recommendations (Rec. 1-17) relating to the four characteristics are considered in identifying the content of the program.

Possible content topics for the teacher education program in classroom assessment are outlined below. These topics reflect the seven assessment competence standards identified for teachers in the United States (American Federation of Teachers, 1990), but do not parallel them. The topics are not intended to be exhaustive, not as an outline for a course. Rather, they serve to give indication of what might be addressed in teacher initial instruction in classroom assessment.

1. Structures of learning outcomes; taxonomies of knowledge, skills, and objectives (discipline-based structures, Bloom's taxonomy, process skills, skill hierarchies, higher-order thinking skills, student-oriented objectives, etc.).

2. Assessment procedures appropriate to particular purposes (decision-making, type of information required, etc.).

3. Principles and skills of paper-and-pencil test construction and item writing (designing a test, creating good items in a variety of formats, marking and scoring procedures, etc.).

4. Principles and skills of accurate and efficient observation (identifying important behaviours, setting systematic observation schemes, interpreting and using observational data, describing observations clearly and efficiently, etc.).

5. Principles and skills of applied performance assessment, including observation (setting appropriate tasks, identifying the behaviours of importance, constructing clear and effective marking schemes, interpreting the information, etc.).

6. Procedures for providing feedback to students and parents (simple diagnostic techniques, oral and written feedback, etc.).

7. Procedures for grading and reporting performance (preparing marks and grades, accurate and concise reporting in a variety of formats).

8. Awareness and sensitivity to issues of fairness, objectivity, personal and other forms of bias, parental and societal expectations.

9. Recognizing and correcting bias reflected in stereotypical portrayal, language choice and structure, and interpretations of patterns of behaviour and response styles.

**Possible constraints and contingencies**. There are a number of features of initial teacher education which are enduring and would be difficult to modify, these would be constraints in adopting any model of teacher preparation in classroom assessment and should be considered in its identification. The constraints would include features of the teacher preparation programs and systems (e.g., faculties of education in universities), and features of the school systems. Constraints inherent in teacher education programs include:

1. The amount of time in the teacher education program which can be devoted to preparation in classroom assessment.

2. The structure of teacher education programs, which usually involves a number of discrete courses that are treated as units based on the amount of student contact time. The programs also include clinical experience, either laboratory or in the school.

3. The variety of educational backgrounds of students in teacher education programs, age levels, academic backgrounds, and experiences.

4. The variety of teaching situations for which teacher education is expected to prepare teachers (e.g., adult education, special needs children).

There are also a number of contingencies in teacher education programs which could be modified if this is deemed necessary:

5. Structuring units of instruction into courses, mini-courses, laboratory work, clinical work, and/or school-based experience. Changes may have implications for restructuring the program, but also for facilities and staffing.

6. Structuring instruction within a discipline, such as Educational Psychology. Changes may require integration of instruction with other aspects of teacher education (e.g., curriculum preparation, pedagogy) and sharing of responsibility. Changes may have implications for organizational responsibilities of the program.

7. Standards of certification. Changes may require certification modifications, such as minimum requirements in classroom assessment preparation, or particular requirements for different teaching specialities.

## Procedures to Review the Recommendations for Teacher Preparation in Classroom Assessment

The 17 recommendations reflect areas of concern within the four categories of reliability, validity, utility, efficiency. Each recommendation should be reviewed with respect to a number of features. To make the review more systematic I ask that you use the scale and categories provided for that feature, and to comment accordingly. Please also make other comments, as you see fit.

**Consider each recommendation for teacher education in classroom assessment.**

- **Importance:** The recommendations are not of equal significance, nor is it expected that they would be given similar amounts of effort in a teacher education program. Your view regarding the importance of the recommendation should be noted, and you may wish to comment on the relative amount of emphasis that should be given to it in teacher preparation.

Check whether you think it is "Very important" or "Not at all important" or somewhere in between.

The four features below are described on pages 15-16.

1. **Level of Specificity:** Amount of substantive and theoretical development, and level of specific, technical prescriptiveness of the instruction in a teacher education program.

Check whether you think it should be "Theoretically based and developed from general principles" or "Specific and providing classroom oriented prescriptions", or somewhere in between.

2. **Common or Differentiated:** Type and level of differentiation of the material and its delivery in a teacher education program, considering such characteristics as specialization.

Check whether you think it should be "Common for all students (prospective teachers)" or "Differentiated for particular groups", or somewhere in between.

3. **Method of Delivery:** Method and timing of delivery of the instruction, and relationship to the overall teacher education program.

Check whether you think it should be "Part of one course in measurement", "Short course", "Seminar", or "Part of a course in curriculum or pedagogy", or some combination of these.

4. **Nature of Instruction:** Nature of the learning experiences for students in a teacher education program.

Check whether you think it should be "Lecture", "Laboratory", "Clinical practice", or "Practical, in-school experience", or some combination of these.

## RELIABILITY: Review of Recommendations

**Recommendation 1. The importance of reliability for high-stakes assessments.** The importance of reliability is directly dependent on the consequences of the assessment. Therefore, teachers must be able to identify the main purpose of an assessment and its implications for the student, both the personal and social consequences (such as in high-stakes assessments).

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## RELIABILITY: Review of Recommendations

**Recommendation 2. Practical ways of improving reliability in the classroom.** The reliability of assessment information can be enhanced in two general ways which are practical in the classroom: by making the assessment procedures more explicit and systematic, and by increasing the amount of high-quality information gathered. Teachers must know how these apply to particular classroom assessment practices.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## RELIABILITY: Review of Recommendations

**Recommendation 3. Increasing reliability through multiple quality assessments.** Students in schools must be given the opportunity to exhibit their skills on several occasions, particularly if there is some doubt if the skill is mastered. Teachers must be able to obtain assessments of skills and knowledge in several different ways.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## RELIABILITY: Review of Recommendations

**Recommendation 4. Numerical procedures for estimating reliability in classroom assessments.** Teachers ought to confirm the adequacy of their measurements and check the reliability of some of the more critical assessments using straightforward numerical procedures.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **1.** Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **2.** Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **3.** Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **4.** Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## RELIABILITY: Review of Recommendations

**Recommendation 5. Reliability of subjective forms of assessment.** Teachers must understand that subjective forms of assessment are prone to unreliability and to personal bias. They must also know practical procedures which can ameliorate these problems.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **1.** Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **2.** Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **3.** Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **4.** Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 6. Increasing assessment of important, higher level cognitive learnings.** Classroom assessments must provide information on the learning that is most important for our students, which includes more complex and higher-level thinking. Therefore, teachers must have the knowledge and skills to design and conduct classroom assessments that are most appropriate to this kind of outcome.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 7. Assessment that supports and informs good instruction.** Teachers must have training and experience in preparing and using assessment procedures that support the most important learning in classrooms, those that should be emphasized in class and on the assessments. Teachers must know how to develop and use paper-and-pencil techniques, but more importantly, they should include systematic observation and applied performance assessment.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 8. Assessment for various purposes including instructional diagnostics.** Teachers must have the training to devise assessment procedures for a variety of purposes, such as for grading of students, individual and group instructional diagnosis, and evaluation of instruction. Teachers must also know when and how it is possible to accommodate more than one purpose, and to recognize when this leads to incompatible procedures.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 9. Empirical checks of validity of the skills covered in classroom assessments.** It is important for classroom assessment that several tasks or items devised to assess the same skill or concept yield similar results with students. Teachers must know how to check for homogeneity. This involves identifying the assessment tasks or items which should "hang together" and checking student results to see if this is the case.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 10. Making assessment results reflect the focus and purpose of the assessment.** Teachers must be trained to obtain scores from assessments corresponding to the nature and purpose of the assessment. This involves identifying clearly in advance of producing the assessment the approximate importance, and hence the weighting, of various areas of content and levels of skills (essentially producing a table of specifications for the assessment).

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 11. Confirming the results of assessments by comparison with other assessments.** Previous assessments and other relevant evaluative information provide a check on assessment results, thereby giving external evidence for their validity. Teachers should understand the importance of external validation, and attempt to obtain several assessments of students' learning using different assessment procedures, and compare performances on these to see if there are glaring discrepancies.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **1.** Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **2.** Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **3.** Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **4.** Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## VALIDITY: Review of Recommendations

**Recommendation 12. Removing bias and prejudice in classroom assessments.** Direct deleterious effects on the assessment results of particular groups of students due to their differences in experience which makes them perform below their actual skill level is an issue of fairness or equity. The kind of bias or prejudice that has a negative reflection on particular subgroups may over the long run have a negative impact on these subgroups. Teachers must recognize and control biases of both types in the materials they use, expectations they have of students, and how they relate to students.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## UTILITY: Review of Recommendations

**Recommendation 13. Making value interpretations of assessment results using referencing procedures.** The interpretation of scores, or other assessment information, is value-laden. Scores are interpreted in reference to norms, criteria, or the individual (self). Teachers must understand the issue of score referencing, and be able to apply the relevant techniques in specific situations. Teachers must also understand that part of the task is setting standards, which involves professional judgement.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## UTILITY: Review of Recommendations

**Recommendation 14. Obtaining the scores necessary to use the assessment results properly.** Appropriate kinds of assessment results and interpretations must be obtainable from the assessment. Teachers must learn how to design the assessment so that the appropriate information is forthcoming. The assessment must produce discrimination of scores at useful points on a scale if the scale is to be converted into grades or mastery-nonmastery categories.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

________________________________________

________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

________________________________________

________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

________________________________________

________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

________________________________________

________________________________________

## UTILITY: Review of Recommendations

**Recommendation 15. Communicating the results and interpretations of assessments.** Various audiences, with different expectations and different abilities, must understand the results. It is necessary to simplify interpretations of assessment results in many cases. Teachers must be able to communicate assessment results in variety of ways and to different audiences. This would include numerical results, summary statistics, grades, anecdotal reports, diagnoses, and affective evaluations.

| **Importance** | Very important | | | Not at all important |
|---|---|---|---|---|
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

## UTILITY: Review of Recommendations

**Recommendation 16. Controlling the effects of subjectivity in observing and marking.** Both objective and subjective forms of assessment are necessary in classroom assessment, so teachers must be skilled in both. Teachers must be trained to use a number of different procedures for subjective evaluations, such as rating schemes, holistic scoring, and narrative descriptions. These procedures can be grounded in subject areas, where particular approaches may be prominent.

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **1.** Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **2.** Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **3.** Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

____________________________________________

____________________________________________

| | | | | | |
|---|---|---|---|---|---|
| **4.** Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

____________________________________________

____________________________________________

## EFFICIENCY: Review of Recommendations

**Recommendation 17. Making maximum use of assessment time and effort.** The amount of assessment effort on the part of the teacher and of the students should be roughly proportional to the significance of the evaluation and the consequences of the decision. Teachers must know ways to speed up the assessment procedures for applied performance assessment including observation, selection-type tests (e.g., multiple choice), and longer constructed response assessments (e.g.,written papers, research reports).

| | | | | |
|---|---|---|---|---|
| **Importance** | Very important | | | Not at all important |
| (check) | ______ | ______ | ______ | ______ |

Comment ______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 1. Level of Specificity | Theoretical basis & general principles | | | Specific, classroom oriented prescriptions | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What theoretical bases? or What kinds of principles or prescriptions?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 2. Common or Differentiated | Common program for all students | | | Differentiated for particular groups | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (If differentiated, on what basis? and What subgroups of students?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 3. Method of Delivery | Part of one course | Short or minicourse | Seminar or lecture | Part of course on pedagogy | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (How should learning be structured? and Where should it be in the program?):

______________________________________________

______________________________________________

| | | | | | |
|---|---|---|---|---|---|
| 4. Nature of Instruction | Lecture/ discussion | Laboratory work | Clinical practice | Practical, in-school experience | Not applicable |
| (check) | ______ | ______ | ______ | ______ | ______ |

Comments (What learning activities and settings should students be involved in?)

______________________________________________

______________________________________________

# APPENDIX E

## RESULTS OF THE REVIEW OF THE 17 RECOMMENDATIONS

Ratings of Importance of the Recommendations .......... 250
Feature 1--Level of Specificity for Instruction .......... 252
Feature 2--Common or Differentiated Instruction .......... 253
Feature 3--Method of Delivery of Instruction .......... 255
Feature 4--Nature of Instruction .......... 257

**Tables**

1. Reviewers' Ratings of Importance of the 17 Recommendations .......... 250
2. Significance of Differences Between Means for Reviewers' Ratings of Importance of the 17 Recommendations .......... 251
3. Reviewer Responses to Level of Specificity of Instruction They Thought Appropriate for Each Recommendation .......... 252
4. Reviewer Responses to Whether Instruction Should be Common or Differentiated for Each Recommendation .......... 254
5. Reviewer Responses to the Method of Instructional Delivery for Each Recommendation .......... 256
6. Reviewer Responses as to Nature of the Instruction for Each Recommendation .......... 257

**Figures**

1. Recommendations Ordered According to Reviewers' Mean Ratings of Importance .......... 251
2. Numbers of Reviewer Responses to Levels of Specificity for Instruction in Assessment .......... 253
3. Numbers of Reviewer Responses to Level of Program Differentiation for Instruction in Assessment .......... 255

## Ratings of Importance of the Recommendations

Reviewers' ratings of the importance of the recommendations were scaled from 4 for *very important* to 1 for *not at all important*. The results for recommendations 1 to 7 were based on 14 reviewers and for recommendations 8 to 17 on 13 reviewers. The means are given in Table 1. Comparisons yielded 136 statistical tests of differences between means on pairs of recommendations. These tests are not independent, but the numbers were too small to apply multivariate statistics. To identify what difference might be considered statistically significant if a pair of means was taken in isolation, a two-tailed *t* test for correlated samples was used. Rather than test all possible pairs directly, average scale variances and covariances were used to obtain an estimate of the size of difference which would be significant. The standard deviations of ratings ranged from .28 (recommendation 7) to .99 (recommendation 17). The standard error of difference in means was estimated by first finding the mean of variances for the 17 recommendations, which was .54, yielding a standard deviation of .74. This mean variance was then adjusted for covariance between scales estimated from the mean for all 136 correlations. The correlations ranged from -.39 to .81, with a mean of .196 (in all, 37 of the 136 correlations were less than 0). Based on this, a difference between means of .544 was required for significance at the .05 level, and .762 for the .01 level.

**Table 1.** Reviewers' Ratings of Importance of the 17 Recommendations (*n* = 14[a]): Scale Ranges from *Very important* (4) to *Not at all important* (1)

| Recommendation | | Mean rating of importance | Recommendation | | Mean rating of importance |
|---|---|---|---|---|---|
| Reliability: | 1. | 3.62 | (validity | 10. | 3.57 |
| | 2. | 3.62 | cont'd) | 11. | 3.00 |
| | 3. | 3.31 | | 12. | 3.54 |
| | 4. | 2.62 | | | |
| | 5. | 3.46 | Utility: | 13. | 3.36 |
| | | | | 14. | 3.50 |
| Validity: | 6. | 3.69 | | 15. | 3.69 |
| | 7. | 3.92 | | 16. | 3.39 |
| | 8. | 3.15 | | | |
| | 9. | 3.00 | Efficiency: | 17. | 3.29 |

[a]One reviewer responded to only seven of the recommendations: *n* = 14 for these recommendations and 13 for the others.

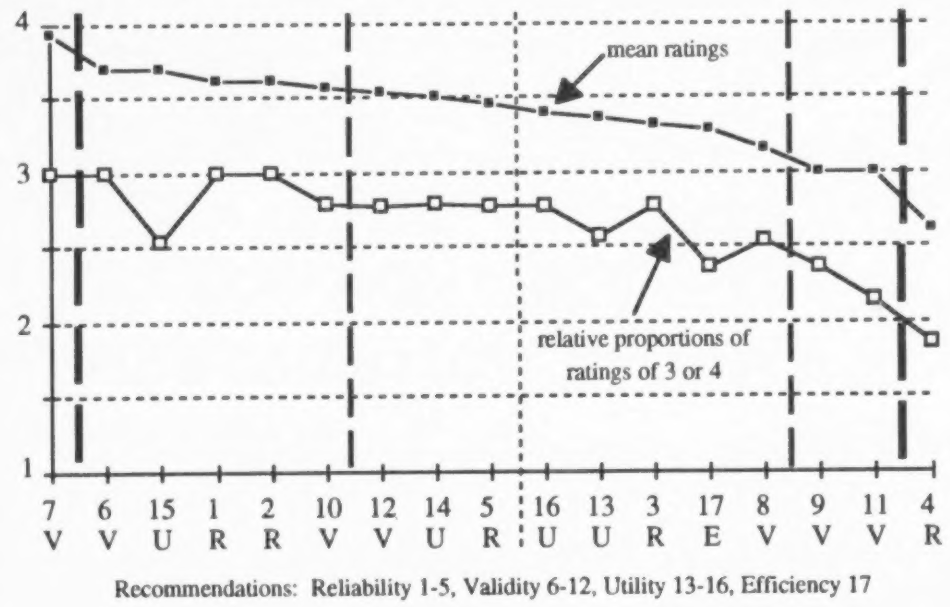
The results of the statistical analyses are reported in Table 2: 17 comparisons were significant at the .05 level and 12 at the .01 level. Recommendation 7 received the highest mean and was rated as significantly more important than seven of the others. Recommendation 4 received the lowest mean and was rated as significantly less important than 13 of the others. Recommendations 6, 15, 1, 2, and 10 were rated higher than recommendation 4 ($p<.01$), and than 9 and 11 ($p<.05$). The recommendations were ordered and grouped according to the mean ratings of recommendations, and this is depicted in Figure 1 by broken vertical lines. The heavier broken lines separate the most clearly differentiated recommendations, 7 and 4. The eight recommendations with mean ratings of 3.4 or lower were separated by a lighter dotted vertical line. The higher rated recommendations include reliability recommendations 1, 2, and 5; validity recommendations 6, 7, 10, and 12; and utility recommendations 14 and 15.

**Table 2.** Significance of Differences Between Means for Reviewers' Ratings of Importance of the 17 Recommendations (correlated *t* test, $df = 12^{a}$)

| Recommendation with **higher** mean | | Recommendations with significantly **lower** means | |
|---|---|---|---|
| | | $p < .05$ | $p < .01$ |
| Reliability: | 1 | 9, 11 | 4 |
| | 2 | 9, 11 | 4 |
| | 3 | 4 | |
| | (4) | | |
| | 5 | | 4 |
| Validity: | 6 | 9, 11 | 4 |
| | 7 | 3, 13, 17 | 4, 8, 9, 11 |
| | (8) | | |
| | (9) | | |
| | 10 | 9, 11 | 4 |
| | (11) | | |
| | 12 | | 4 |
| Utility: | 13 | 4 | |
| | 14 | | 4 |
| | 15 | 9, 11 | 4 |
| | 16 | | 4 |
| Efficiency: | 17 | 4 | |

[a] One reviewer responded to only seven of the recommendations, so for most pairs of scales there were 13 usable responses.

**Figure 1.** Recommendations Ordered According to Reviewers' Mean Ratings of Importance (scale of 4 to 1): Also Shows Proportions of Reviewers Rating Each Recommendation as *Very important* (4) or the Next Scale Category (3)



Note. The proportions of reviewers selecting *very important* (4) or next category on the scale (3) were converted to a scale from 0%, indicated by the scale point 1, to 100%, indicated by the scale point 3, so that these could be displayed on the same chart as the means.

The number of ratings of 4 and 3 were combined for each recommendation and converted to a relative proportion scale ranging from 1 (0%) to 3 (100%) so that they could be depicted on Figure 1 along with the means. This graph roughly parallels that of the means, with clear exceptions for recommendations 15, 13, and 17. Using this graph it is also more difficult to differentiate recommendations 15, 10, 12, 14, and 5 from recommendations 16 and 3. However, recommendations 13, 17, 8, 9, 11, and 4 appear to be rated lower than those above the dotted line, except for recommendation 15. The means take into account the scale difference between 4 and 3 and between 1 and 2, so these were used to determine the order of importance of the recommendations.

## Feature 1--Level of Specificity for Instruction

Responses to the level of specificity of instruction were scaled from 4 for *theoretically based and developed from general principles* to 1 for *specific and providing classroom-oriented prescriptions.* Some reviewers checked two or more categories on the scale, but, for the most part, reviewers gave only one response for each recommendation. This can be seen from the *N of responses* column in Table 3 below, and only for recommendation 15 were there more than four responses beyond one per reviewer. For multiple responses the mean scale value of these responses was entered. Means for the 17 recommendations are reported in Table 3. The scale range is from 1 to 4 so a theoretical middle point is 2.5. A mean above this point suggests that the instruction for the recommendation should be more theoretically based, and a mean below 2.5 suggests instruction should give specific classroom prescriptions. Means were tested to determine if they deviated significantly from 2.5 using a two-tailed *t* test. This does not account for the fact that the tests form a non-independent set whose combined significance level is far from the nominal one for each test, but does give some guidance in interpretation of the results. The mean for recommendation 2 was significantly below 2.5, and the means for recommendations 8, 9, 11, and 13 were significantly above 2.5.

**Table 3**. Reviewer Responses to Level of Specificity of Instruction They Thought Appropriate for Each Recommendation ($n = 15$)

Feature 1. Level of specificity: scaled from *theoretical basis* (4) to *specific prescriptions* (1), and compared to a theoretical middle point of 2.5 using a two-tailed *t* test

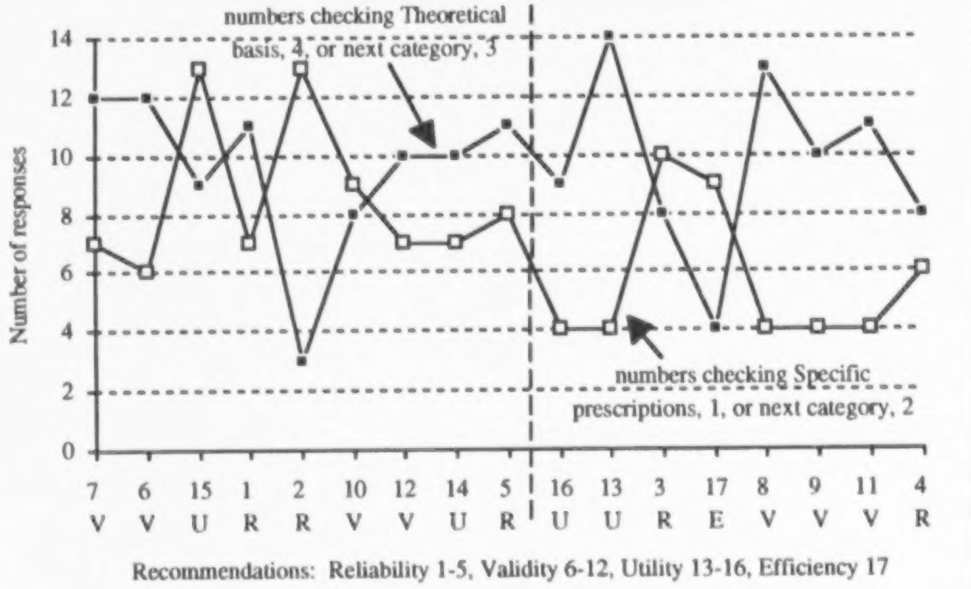
| Recommendation | | *N* of responses[a] | Rating | | Recommendation | | *N* of responses[a] | Rating | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | $p <$ | | | | Mean | $p <$ |
| Reliability | 1. | 18 | 2.7 | ns | (validity | 10. | 17 | 2.5 | ns |
| | **2.** | **16** | **1.9** | **.05** | cont'd) | **11.** | **15** | **3.3** | **.05** |
| | 3. | 19 | 2.3 | ns | | 12. | 17 | 2.8 | ns |
| | 4. | 17 | 2.9 | ns | | | | | |
| | 5. | 19 | 2.7 | ns | Utility | **13.** | **18** | **3.3** | **.01** |
| | | | | | | 14. | 17 | 2.7 | ns |
| Validity | 6. | 18 | 2.8 | ns | | 15. | 22 | 2.2 | ns |
| | 7. | 19 | 2.8 | ns | | 16. | 14 | 2.8 | ns |
| | **8.** | **17** | **3.0** | **.05** | | | | | |
| | **9.** | **15** | **3.2** | **.01** | Efficiency | 17. | 15 | 2.3 | ns |

[a] *N* can exceed 15 since some reviewers checked more than one response category. The mean response was taken for these individuals.

For the second analysis the numbers of responses were combined for scale points 4 and 3, and for 2 and 1, to determine whether instruction should be primarily theoretically based (4 and 3) or emphasize specific classroom prescriptions (2 and 1). These results are depicted as a graph in Figure 2. For recommendations 6 and 7 most of the reviewers thought instruction could be more theoretically based (12 responses), although the means were not significantly above 2.5. For recommendations 15 and 2, 13 responses indicated instruction should be specific and prescriptive, and for recommendation 2 the mean was significantly above 2.5. For recommendation 1 more reviewers (11 responses) thought instruction should be theoretical than specific prescriptions (7 responses), although the mean was not significantly different from 2.5. For recommendation 10 reviewers were fairly evenly split as to whether instruction should be theoretical or specific. For recommendations 12, 14, and 5 more reviewers (10-11 responses) selected theoretical instruction, although the means did not differ significantly from 2.5.

For six of the eight recommendations rated less important, more reviewers (8-14 responses) thought instruction should be theoretically based: recommendations 16, 13, 8, 9, 11, and 4. The means for recommendations 13, 8, 9, and 11 were significantly higher than 2.5. For recommendations 3 and 17 specific instruction was favoured by more reviewers (9-10 responses), but the means did not differ significantly from 2.5.

**Figure 2.** Numbers of Reviewer Responses to Levels of Specificity for Instruction in Assessment, on the Scale *Theoretical basis* (4) to *Specific prescriptions* (1) ($n = 15$)



Note. The numbers of reviewer responses selecting *theoretical basis* (4) or next scale category (3) were combined, and numbers selecting *specific prescriptions* (1) or next scale category (2) were combined.

## Feature 2--Common or Differentiated Instruction

Responses to whether the instruction should be differentiated for certain subgroups of students were scaled from 4 for *common program for all students* to 1 for *differentiated for particular groups*. Some reviewers checked two or more categories on the scale, but, for the most part, reviewers gave only one response for each recommendation (see the *N of responses* column in Table 4). For multiple responses the mean scale value of these responses was entered. As with Feature 1, means were analyzed using a two-tailed $t$ test to determine generally if it was thought that instruction on a recommendation should be common to all students, mean above 2.5, or differentiated, mean below 2.5 (Table 4). The 17 tests form a non-independent set whose combined significance level is above .05, but they do give some guidance for the interpretation.

**Table 4**. Reviewer Responses to Whether Instruction Should be Common or Differentiated for Each Recommendation ($n = 15$)

Feature 2. Common or differentiated instruction: scaled *common* (4) to *differentiated* (1), and compared to a theoretical middle point of 2.5 using a two-tailed $t$ test

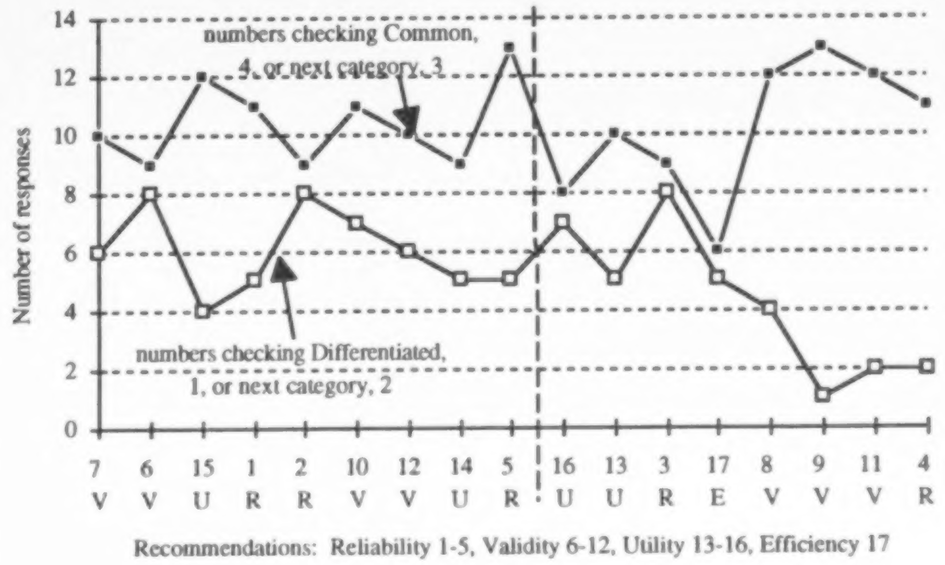
| Recommendation | | $N$ of responses[a] | Rating | | Recommendation | | $N$ of responses[a] | Rating | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | $p <$ | | | | Mean | $p <$ |
| Reliability | 1. | 16 | 3.0 | ns | (validity cont'd) | 10. | 18 | 2.9 | ns |
| | 2. | 17 | 2.6 | ns | | **11.** | **14** | **3.3** | **.05** |
| | 3. | 18 | 2.7 | ns | | 12. | 16 | 3.0 | ns |
| | **4.** | **15** | **3.4** | **.01** | | | | | |
| | **5.** | **18** | **3.2** | **.05** | Utility | 13. | 15 | 3.1 | ns |
| | | | | | | 14. | 15 | 2.9 | ns |
| Validity | 6. | 17 | 2.7 | ns | | **15.** | **17** | **3.3** | **.01** |
| | 7. | 16 | 2.9 | ns | | 16. | 16 | 2.6 | ns |
| | **8.** | **16** | **3.0** | **.05** | | | | | |
| | **9.** | **15** | **3.6** | **.01** | Efficiency | 17. | 14 | 2.7 | ns |

[a] $N$ can exceed 15 since some reviewers checked more than one response category. The mean response was taken for these individuals.

Means were significantly above 2.5, $p<.01$, for recommendations 4, 9, and 15, and $p<.05$, for recommendations 5, 8, and 11, suggesting that instruction for these recommendations should be generally common for students. No mean was below 2.5. This indicates for the remaining recommendations, that although reviewers were not definitive as to whether instruction should be common or differentiated, generally common instruction appeared to be more favoured.

Numbers of responses were then combined for scale points 4 and 3, reflecting common instruction, and for 2 and 1, reflecting differentiated instruction. The recommendations were ordered according to their importance, and the combined frequencies were depicted as a graph in Figure 3 with a dotted line separating the nine recommendations rated higher in importance from the others. Only slightly more reviewers chose common instruction as chose differentiated instruction for recommendations 2, 6, and 7, and instruction could be either common or differentiated. More reviewers, 12 to 13, favoured common instruction over differentiated for recommendations 5 and 15, and the means were significantly above 2.5. Means were not significantly different from 2.5 for the other recommendations that were rated more important, but reviewers appeared to favour common instruction for recommendations 1, 10, 12, and 14, with from 9 to 11 responses in categories 3 or 4, versus from 5 to 7 in categories 1 or 2. Four of the recommendations rated lower in importance obtained means significantly above 2.5, recommendations 8, 9, 11, and 4. From 11 to 13 responses were in categories 3 or 4, clearly supporting common instruction. The remaining recommendations (16, 13, 3, 17), whose means were not significantly different from 2.5, obtained from 5 to 8 responses in categories 1 or 2, and almost as many as responses in 3 or 4, which were from 8 to 10.

**Figure 3.** Numbers of Reviewer Responses to Level of Program Differentiation for Instruction in Assessment, on the Scale *Common for all students* (4) to *Differentiated for particular groups* (1)



Note. The numbers of reviewer responses selecting *common program* (4) or next scale category (3) were combined, and numbers selecting *differentiated* (1) or next scale category (2) were combined.

## Feature 3--Method of Delivery of Instruction

Reviewers indicated the method of delivery of instruction for each recommendation by selecting one or more of four categories: *part of one course*, *short or minicourse*, *seminar or lecture*, and *part of course on pedagogy*. There was no meaningful scale underlying the response categories, so the frequencies of responses were compared across the categories for each recommendation using a $\chi^2$ goodness-of-fit test based on the null hypothesis of equal frequencies of responses for each category (25% of responses). These analyses are reported in Table 5. As with the tests for the first two features, this involved non-independent tests, and the number of respondents was small. Expected frequencies approached 5, which is the minimum often suggested for goodness-of-fit tests (e.g., Ferguson & Takane, 1989; Shavelson, 1988). Further, some of the responses to one recommendation clearly were not independent since often one reviewer selected more than one category. Nevertheless, these analyses were used to help guard against over interpretation of perceived differences.

Significance at the .05 level was obtained for four of the recommendations: 5, 6, 7, and 8, of which the first three were rated as more important by the reviewers ; these recommendations are highlighted by boldface type in Table 5. The frequencies for the first and last category were larger for recommendations 5 and 7, suggesting more support for delivery as *part of one course* and *part of a course on pedagogy*, the latter receiving by far the highest frequencies of response. For recommendations 6 and 8 the same two categories were favoured, but with nearly equal frequencies. Few reviewers thought short courses or seminars were appropriate for any of these four recommendations.

**Table 5.** Reviewer Responses to the Method of Instructional Delivery for Each Recommendation ($n = 15$)

Feature 3. Method of delivery of the instruction: number of responses to the four categories, compared to a theoretical expected proportion of 25% using a $\chi^2$ test

| Recommendation | Not applicable | Part of one course | Short or mini-course | Seminar or lecture | Part of course on pedagogy | $N$ of usable responses[a] | $\chi^2$ ($df = 3$) | $p <$ |
|---|---|---|---|---|---|---|---|---|
| Rel. 1 | – | 8 | 1 | 3 | 7 | 19 | 6.89 | ns |
| 2 | – | 8 | 2 | 3 | 6 | 19 | 4.79 | ns |
| 3 | 1 | 5 | 2 | 4 | 8 | 19 | 3.95 | ns |
| 4 | 2 | 6 | 5 | 3 | 3 | 17 | 1.59 | ns |
| **5** | -- | **6** | **2** | **2** | **11** | **21** | **10.43** | **.05** |
| Val. **6** | -- | **7** | **0** | **2** | **8** | **17** | **10.53** | **.05** |
| **7** | -- | **6** | **2** | **3** | **12** | **23** | **10.57** | **.05** |
| **8** | -- | **8** | **1** | **1** | **6** | **16** | **9.50** | **.05** |
| 9 | 1 | 8 | 1 | 3 | 4 | 16 | 6.50 | ns |
| 10 | – | 5 | 5 | 2 | 7 | 19 | 2.68 | ns |
| 11 | – | 7 | 4 | 2 | 3 | 16 | 3.50 | ns |
| 12 | – | 4 | 4 | 5 | 9 | 22 | 3.09 | ns |
| Ut. 13 | – | 7 | 3 | 3 | 8 | 21 | 3.95 | ns |
| 14 | – | 6 | 5 | 2 | 5 | 18 | 2.00 | ns |
| 15 | – | 7 | 2 | 2 | 8 | 19 | 6.47 | ns |
| 16 | – | 5 | 3 | 4 | 11 | 23 | 6.74 | ns |
| Eff. 17 | 2 | 3 | 2 | 2 | 8 | 15 | 6.60 | ns |
| **Totals** | **6** | **106** | **44** | **46** | **124** | **320** | **63.30** | **.01** |

[a] Some reviewers checked more than one category: $N$ is the number of usable responses.

There appeared to be a general finding that the categories of part of a course and part of pedagogy were favoured for all recommendations: the two categories received total frequencies of 106 and 124 as opposed to the short courses and seminars receiving 44 and 46 responses respectively. However, there were recommendations for which the reviewers did not favour one method over the others.

There the other recommendations that were rated as more important, 1, 2, 10, 12, 14, and 15, it was not as clear what method of delivery the reviewers favoured, although for recommendations 1, 2, and 15 the two categories, part of one course and part of a course on pedagogy, were endorsed with almost equal frequency (6 to 8 responses) and more frequently than the other two categories by 3 to 6 responses. For recommendation 12 the category of part of a course on pedagogy was selected by slightly more reviewers, but the frequencies for the other three categories were similar. Recommendations 10 and 14 received similar frequencies for three categories, with the seminar category receiving somewhat fewer responses.

The remaining seven recommendations received lower ratings of importance. For recommendations 3, 16, and 17 the category, part of a course on pedagogy, received from 3 to 6 more responses than did the other three categories. For recommendations 9 and 11 the category, part of one course, received from 3 to 4 more responses than did the other three categories. For recommendation 13 the two categories, part of one course and part of a course on pedagogy, received similar numbers of responses and more than the other two categories. Recommendation 4 received similar numbers of responses for all four categories.

## Feature 4--Nature of Instruction

Reviewers indicated the nature of the instruction for each recommendation by selecting one or more of four categories: *lecture/discussion*, *laboratory work*, *clinical practice*, and *practical in-school experience*. As with Feature 3, there was no meaningful scale underlying the response categories, so the frequencies of responses were compared for each recommendation using a $\chi^2$ goodness-of-fit test based on the null hypothesis of equal frequencies of responses for each category (25% of responses). These are reported in Table 6. There were problems with these statistical tests, as there were for those conducted for features reported on earlier: the tests were not independent, the number of respondents was small, expected frequencies approached 5, and some reviewers selected more than one category. However, again these analyses were used to guard against over interpretation of minor differences.

**Table 6**. Reviewer Responses as to Nature of the Instruction for Each Recommendation ($n = 15$)

Feature 4. Nature of instruction: number of responses to the four categories, compared to a theoretical expected proportion of 25% using a $\chi^2$ test

| Recom-mendation | | Not applic-able | Lecture/ dis-cussion | Lab-oratory work | Clinical practice | Practical in-school exper. | *N* of usable responses[a] | $\chi^2$ ($df = 3$) | $p <$ |
|---|---|---|---|---|---|---|---|---|---|
| Rel. | 1 | – | 12 | 7 | 6 | 6 | 31 | 3.19 | ns |
| | 2 | – | 10 | 4 | 8 | 7 | 29 | 2.59 | ns |
| | 3 | – | 6 | 7 | 8 | 8 | 29 | 0.38 | ns |
| | **4** | **2** | **10** | **9** | **2** | **1** | **22** | **11.82** | **.01** |
| | 5 | – | 11 | 5 | 6 | 3 | 25 | 5.56 | ns |
| Val. | 6 | – | 7 | 4 | 8 | 7 | 26 | 1.38 | ns |
| | 7 | – | 8 | 6 | 8 | 10 | 32 | 1.00 | ns |
| | **8** | **–** | **11** | **6** | **0** | **4** | **21** | **11.95** | **.01** |
| | 9 | 1 | 10 | 5 | 4 | 3 | 22 | 5.27 | ns |
| | 10 | – | 10 | 3 | 5 | 5 | 23 | 4.65 | ns |
| | 11 | – | 8 | 6 | 4 | 4 | 22 | 2.00 | ns |
| | 12 | 1 | 11 | 6 | 6 | 5 | 28 | 3.14 | ns |
| Ut. | 13 | – | 12 | 5 | 3 | 6 | 26 | 6.92 | ns |
| | 14 | – | 10 | 7 | 3 | 4 | 24 | 5.00 | ns |
| | 15 | – | 12 | 4 | 3 | 7 | 26 | 7.54 | ns |
| | 16 | – | 7 | 4 | 6 | 6 | 23 | 0.83 | ns |
| Eff. | 17 | 2 | 6 | 6 | 3 | 4 | 19 | 1.42 | ns |
| **Totals** | | **6** | **161** | **94** | **83** | **90** | **428** | **36.92** | **.01** |

[a] Some reviewers checked more than one category: *N* is the number of usable responses.

Recommendations 4 and 8 were significant, and at the .01 level, indicating differing numbers of responses for various categories. These are highlighted by boldface type in Table 6. Reviewers clearly favoured the categories lecture/discussion (10-11 responses) and laboratory work (6-9 responses). When totaled for all 17 recommendations, there were 161 responses favouring the lecture and discussion category and only 80 to 90 favouring the other three (significant at $p<.01$). The first category received 25% or more of the responses for all recommendations other than recommendation 3, suggesting that the reviewers generally supported instruction using lecture and discussion for the recommendations, and that the other three approaches were considered less appropriate depending on the recommendation.

However, significance was not obtained for any of the nine recommendations rated higher in importance by the reviewers. For recommendations 6 and 7 frequencies were very similar support for all four categories of instruction. There appeared to be more responses favouring lecture and discussion for recommendations 1, 5, 10, 12, and 15, where frequencies for this category exceeded those for the other three by 5 or more responses. The other three categories received similar numbers of responses for recommendations 1, 5, and 10, and the category in-school experience received 3 to 4 more responses than the other two for recommendation 15. For recommendations 2 and 14 the responses to the lecture and discussion category exceeded those to the other categories by 2 or 3 responses. Clinical experience and in-school experience received nearly as many responses for recommendation 2, whereas laboratory work received nearly as many for recommendation 14.

The remaining six recommendations had received lower ratings of importance. For recommendations 3, 16, and 17, the numbers of responses deviated from one another by 3 or less, suggesting that all approaches were considered similarly appropriate. For recommendations 9 and 13 the category, lecture and discussion, received from 5 to 7 more responses than did the other three categories. Finally, for recommendation 11, the two categories lecture and discussion and laboratory work seemed to be supported more than the other two.